Approches géométriques de l'analyse de données génétiques

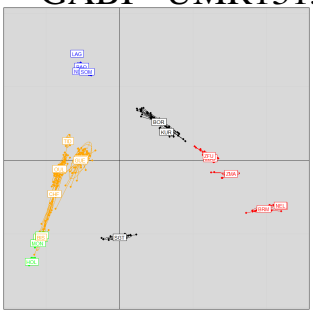
Denis Laloë

INSTITUT NATIONAL POLYTECHNIQUE DE TOULOUSE

HABILITATION À DIRIGER DES RECHERCHES
SOUTENUE LE 15 OCTOBRE 2015

Approches géométriques de l'analyse de données génétiques

Denis LALOË
GABI - UMR1313



MEMBRES DU JURY

Pr David Causeur	AgroCampus Ouest
Dr Alain Charcosset	INRA - Le Moulon
Pr Alain Ducos	Ecole Vétérinaire de Toulouse
Dr Anne-Béatrice Dufour	Université de Lyon
Dr Christèle Robert	INRA-Toulouse

Table des matières

Avant-propos

Titulaire d'un diplôme d'ingénieur agronome de l'ENSA de Rennes, d'un DEA de génétique quantitative et appliquée - Université Paris-Sud, et d'un DEA de modélisation stochastique et statistique - Université Paris-Sud, j'ai été recruté à l'INRA en 1987 comme ingénieur de recherches, à la Station de Génétique Quantitative et Appliquée (département de Génétique Animale), à Jouy-en-Josas. De 1987 à 2010, j'ai construit et conduit l'évaluation génétique en ferme des bovins allaitants (IBOVAL), dans l'équipe « Amélioration Génétique des Bovins Allaitants », puis, lors de la création de la TGU [1] GABI « Génétique Animale et Biologie Intégrative », dans l'équipe G2B (Génétique et Génomique Bovine). En 2011, j'ai rejoint l'équipe PSGen (Populations, Statistique et Génome), comme ingénieur biostatisticien. J'en assure l'animation depuis 2012.

Je ne suis pas un statisticien. Je suis un généticien analyste de données génétiques, *ie* des données que l'on peut relier à des phénomènes de nature génétique (généalogies, marqueurs génétiques). Née avec Mendel, cette génétique phénoméniste s'intéresse seulement à la quantification des phénomènes liés à la génétique (par exemple, des performances d'apparentés ou des fréquences alléliques), et non à la nature physique du matériel génétique (Laloë, 2011).

Une analyse de données est formée d'un couple à marier, le méthodologiste, et le client fournisseur de données, et d'un marieur, l'analyste. A charge de celui-ci d'établir la correspondance entre l'interprétation possible des résultats telle qu'elle ressort de la méthode utilisée, et l'utilisation réelle qui en est projetée par le client. Les critères de jugement sont différents selon les points de vue. Les objectifs et propriétés statistiques d'une méthode donnée ne recoupent pas forcément les objectifs recherchés par l'utilisateur des résultats. Le méthodologiste ignore l'objectif réel de l'utilisateur, alors que celui-ci ignore les contraintes de la méthode.

Mes activités de recherche ont couvert les deux grands domaines de l'analyse de données génétiques (génétique quantitative et génétique des populations), avec quelques incursions dans l'analyse des données zootechniques. Selon ces domaines, mon implication a varié. J'ai été responsable d'une évaluation génétique, avec la maîtrise, autant que faire se peut, du processus d'analyse : gestion des données, choix des modèles et des méthodes, calcul et interprétation des résultats. Au contraire, j'ai analysé des données zootechniques en étant extérieur au sujet de recherche, sans prise sur le recueil des données et réel investissement sur la discussion des résultats des analyses.

1. Très Grande Unité

Génétique Quantitative : J'ai mis en place et conduit l'évaluation génétique des bovins allaitants à partir de leurs performances en ferme de la naissance au sevrage (*IBOVAL*), avec le coencadrement de deux thèses (M J Shi [2] et M N Fouilloux [3]), une participation à la thèse de R Rincent [4] et l'encadrement de 5 masters (ou fins d'études d'ingénieurs). Mes recherches ont porté sur la connexion génétique entre troupeaux, en prenant en compte les spécificités d'une évaluation génétique.

Génétique des populations : J'ai étudié la structuration génétique et de la diversité des populations domestiques à partir de données moléculaires, avec le coencadrement de la thèse de K Moazami-Goudarzi [5] et l'encadrement de deux masters. Collaborateur des projets, mon rôle a consisté à proposer et adapter des méthodes d'analyse factorielle, où la dualité observations/variables spécifique de ces méthodes permet d'appréhender le rôle des marqueurs dans une représentation consensuelle de la structure génétique de populations proches.

Données zootechniques J'ai joué dans ce domaine un rôle d'analyste expert, en proposant l'utilisation des analyses factorielles dans des études zootechniques, en participant à l'encadrement de deux thèses sur l'élevage porcin (L Canario [6] ; B Salmi [7]).

Après une introduction qui précisera en quoi les données génétiques se prêtent à une approche géométrique, un premier chapitre abordera comment l'analyse des corrélations canoniques permet d'appréhender l'ensemble des problématiques liées à la précision dans une évaluation génétique ; un deuxième chapitre traitera des différents avatars de l'analyse factorielle dans l'étude des données de génétique des populations et zootechniques ; le troisième concluera sur mes perspectives de recherche.

2. Meng-Jiao SHI, Estimation of direct and maternal variability of preweaning traits in beef cattle. Application to field data of french beef breeds. Institut National Agronomique Paris-Grignon, département des sciences animales. Soutenue en 1993

3. Marie-Noëlle FOUILLOUX, Amélioration des systèmes d'amélioration génétique des aptitudes bouchères des reproducteurs de races bovines allaitantes. Institut National Agronomique Paris-Grignon, département des sciences animales, soutenue en 2000

4. Renaud RINCENT, Optimisation des stratégies de génétique d'association et de sélection génomique pour des populations de diversité variable. Application au maïs, AgroParisTech, soutenue en 2014

5. Katayoun MOAZAMI-GOUDARZI. Caractérisation de plusieurs races bovines françaises à l'aide de marqueurs polymorphes, université Paris-Sud, soutenue en 1994.

6. Laurianne CANARIO. Aspects génétiques de la mortalité des porcelets à la naissance et en allaitement précoce : relations avec les aptitudes maternelles des truies et la vitalité des porcelets. Institut National Agronomique Paris-Grignon, soutenue en 2006

7. Btissam SALMI. Méta-analyse et analyse multidimensionnelle : applications à la qualité de la viande de porc. AgroParisTech, soutenue en 2011

Introduction

Pour une large part, les statistiques se préoccupent d'extraire de l'information pertinente à partir de masses importantes de données numériques, souvent structurées en tableaux. Il est naturel que les statistiques s'appuient sur une approche géométrique, où la représentation des données se fait sous forme de nuages de points, et sur l'algèbre linéaire : on y parle d'espaces, de distances, de métriques, d'angles, de projection.

Géométrie et modèle linéaire

Des statisticiens aussi éminents que Fisher ou Bartlett ont utilisé cette approche dans le cadre du modèle linéaire (Herr, 1980) et de l'optimisation des plans d'expérience (Silvey and Titterington, 1973). Fisher développa une représentation géométrique dans $\mathbb{R}^n$ des variables statistiques, de leurs moyennes, variances et corrélations, ainsi que de la décomposition des espaces vectoriels en sous-espaces orthogonaux mobilisés en analyse de variance (Armatte, 2008). L'estimateur $\hat{\mathbf{y}}$ des moindres carrés du modèle $\mathbf{y} = \mathbf{X}\beta + e$ est la projection orthogonale du vecteur $\mathbf{y}$ sur le sous-espace de $\mathbb{R}^n$ engendré par les colonnes de $\mathbf{X}$: (cf Figure 1.)

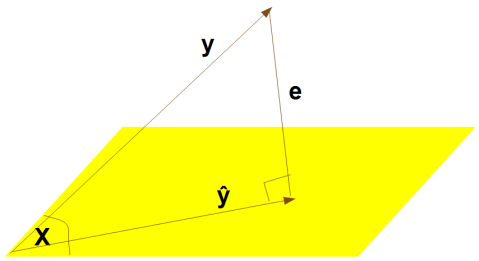


FIGURE 1 – Estimateur des moindres carrés - modèle $\mathbf{y} = \mathbf{X}\beta + e$

Cette approche est détaillée dans Scheffé (1959), pour qui *"One unifying and insightful way of regarding the analysis of variance is from the geometrical viewpoint :it may viewed as a method of resolving the vector of observations into vectors lying in certain spaces corresponding to different sources of variation in the observations, and to each of which a meaningful interpretation can be given"*. Scheffé (1953) considére l'ensemble des contrastes comme un

espace vectoriel. L'ensemble des contrastes estimables est un sous-espace vectoriel du précédent. Comme conséquence, tout contraste n'appartenant à ce sous-espace n'est pas estimable. J'ai repris cette approche pour l'étude de la disconnexion des effets aléatoires en modèle mixte.

Géométrie et analyse factorielle des données

La géométrie est au coeur des analyses factorielles, dès l'article fondant l'analyse en composantes principales de Pearson (1901) : *On lines and planes of closest fit fo systems of points in space*. Le Roux (2014) a d'ailleurs regroupé les analyses factorielles sous le terme générique d'*Analyse Géométrique des Données* . L'analyse factorielle traite un tableau de données où les individus (lignes) sont décrits par des variables (colonnes) sous forme de deux nuages de points, l'un pour les lignes, l'autre pour les colonnes. Chacun des nuages est dans un espace vectoriel muni d'une métrique, de façon à ce que les distances entre points reflètent une proximité statistique. Ces deux nuages sont projetés sur une suite d'axes orthogonaux qui maximisent l'inertie d'une façon unique et optimale, en vertu du théorème d'Eckart and Young (1936). Le point central de l'analyse est la dualité entre lignes et colonnes du tableau, ou la dualité des espaces vectoriels associés, qui permet la représentation et l'interprétation simultanée des deux espaces. Escoufier (1987) a formalisé cette dualité par le schéma de dualité (cf Figure 2) .

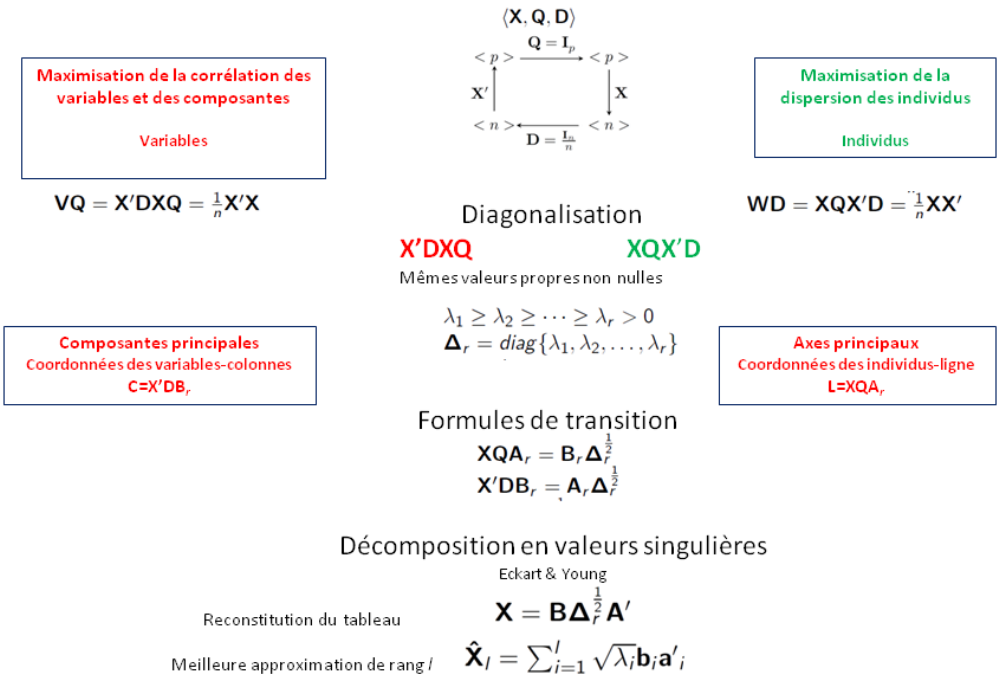


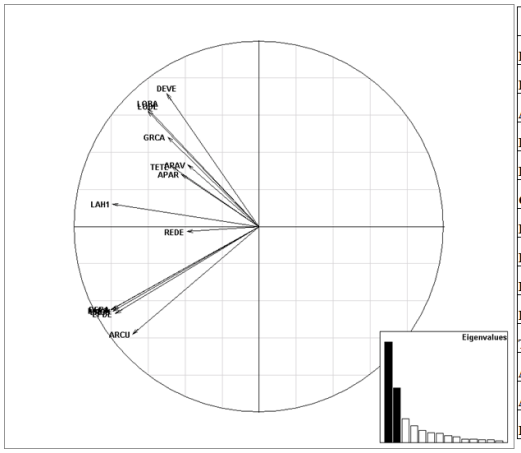
FIGURE 2 – Schéma de dualité. Application à l'Analyse en Composantes Principales

Ces méthodes appréhendent un phénomène dans sa généralité, selon une vision holistique, synthétique, seule à même, pour Benzécri (Armatte, 2008), de faire émerger un système de relations causales. Les données priment sur les modèles, selon la célèbre citation de Benzécri : *"Le modèle doit suivre les données, non l'inverse"*. Enfin, la place essentielle conférée aux graphiques dans la restitution des résultats permet, comme l'écrivent Escofier and Pagès (1998), *"d'utiliser les facultés de perception dont nous usons quotidiennement : sur les graphiques de l'analyse factorielle, on voit, au sens propre du terme (avec les yeux et l'analyse assez mystérieuse que notre cerveau fait d'une image), des regroupements, des oppositions, des tendances, impossibles à discerner directement sur un grand tableau de nombres, même après un examen prolongé. Ces représentations graphiques sont aussi un moyen de communication remarquable car point n'est besoin d'être statisticien pour comprendre que la proximité entre deux points traduit la ressemblance entre les objets qu'ils représentent"*. L'analyse éclaire les données d'une façon différente, plus complète, apportant ainsi une réelle plus-value en terme de connaissance et d'interprétation pour l'utilisateur. Le graphique est un objet commun à l'analyste des données et à l'utilisateur, une mise en forme des données (sens premier du terme *information*), immédiatement perceptible. C'est primordial dans un dialogue avec des utilisateurs qui voient souvent les statistiques comme un mal nécessaire pour obtenir le tampon **SIGNIFICATIF**. Le cercle des corrélations (Figure 3) [8], extraite d'une étude sur la morphologie des bovins allaitants, synthétise l'essentiel :

- un premier axe confirmant un facteur de développement général ;
- un deuxième axe, montrant des différences de forme, opposant le développement des masses musculaires et le développement squelettique ;
- le regroupement des notes selon des aptitudes générales, traduisant la cohérence interne intra-aptitudes.

Ce cercle des corrélations a permis aux techniciens de terrain d'objectiver leur perception d'une opposition entre développements musculaire et squelettique, jusque-là cachée par les corrélations positives entre ces deux aptitudes. Du coup, l'analyste de données a acquis à leurs yeux le statut d'interlocuteur légitime.

8. Les méthodes utilisées et les graphiques présentés dans ce document proviennent tous du package R *ade4* (Dray and Dufour, 2007).



	DEPA	LADO	ARCU	LACU	EPDE	GRCA	LODE	LBA	LAH	DEV	TETE	APV	APR	REDE
DEPA	1	0,89	0,73	0,78	0,78	0,19	0,22	0,21	0,54	0,1	0,23	0,13	0,17	0,23
LADO	0,89	1	0,74	0,78	0,79	0,19	0,21	0,2	0,53	0,09	0,22	0,16	0,19	0,33
ARCU	0,73	0,74	1	0,81	0,8	0,06	0,09	0,07	0,43	-0,04	0,13	0,05	0,14	0,23
LACU	0,78	0,78	0,81	1	0,8	0,17	0,21	0,21	0,6	0,08	0,2	0,12	0,17	0,24
EPDE	0,78	0,79	0,8	0,8	1	0,15	0,2	0,19	0,52	0,08	0,21	0,13	0,17	0,29
GRCA	0,19	0,19	0,06	0,17	0,15	1	0,46	0,5	0,46	0,48	0,42	0,28	0,29	0,07
LODE	0,22	0,21	0,09	0,21	0,2	0,46	1	0,84	0,52	0,82	0,35	0,27	0,27	0,17
LOBA	0,21	0,2	0,07	0,21	0,19	0,5	0,84	1	0,54	0,8	0,37	0,28	0,28	0,16
LAH1	0,54	0,53	0,43	0,6	0,52	0,46	0,52	0,54	1	0,45	0,37	0,27	0,28	0,19
DEVE	0,1	0,09	-0,04	0,08	0,08	0,48	0,82	0,8	0,45	1	0,34	0,28	0,26	0,16
TETE	0,23	0,22	0,13	0,2	0,21	0,42	0,35	0,37	0,37	0,34	1	0,3	0,29	0,12
APAV	0,13	0,16	0,05	0,12	0,13	0,28	0,27	0,28	0,27	0,28	0,3	1	0,5	0,29
APAR	0,17	0,19	0,14	0,17	0,17	0,29	0,27	0,28	0,28	0,26	0,29	0,5	1	0,24
REDE	0,23	0,33	0,23	0,24	0,29	0,07	0,17	0,16	0,19	0,16	0,12	0,29	0,24	1

FIGURE 3 – Morphologie des bovins. A droite, la matrice des corrélations ; à gauche, l'ACP (Cercle des corrélations)

Les données génétiques

Pour séduisantes que soient les approches exposées ci-dessus, elles ne donnent leur pleine mesure que pour des données multidimensionnelles. C'est le cas des données génétiques issues de ses branches quantitative et des populations, qui sont fondées sur les variations entre individus, leurs mesures, leurs causes et leurs conséquences.

Génétique quantitative. Là où la population idéale du statisticien se compose d'individus échangeables, indépendamment et identiquement distribués, qu'il suffit de caractériser par quelques statistiques exhaustives, le généticien quantitatif sait depuis 1903, grâce au botaniste danois Willhem Johannsen (Laloë, 2011), qu'une population est faite d'individus différents les uns des autres, et, qu'à l'inverse, un individu ne peut se réduire à la population à laquelle il appartient.

Génétique des populations. La caractérisation des individus en génétique des populations est aujourd'hui massivement multidimensionnelle. D'une vingtaine de microsatellites il n'y a pas si longtemps (Moazami-Goudarzi et al., 1997), on est passé aujourd'hui à des milliers, voir des millions de marqueurs SNP, sans même parler des séquences.

1

Corrélations canoniques et précision d'une évaluation génétique

1.1 Qu'est-ce qu'une évaluation génétique ?

1.1.1 Les fondamentaux

Fondamentalement, une évaluation génétique répond à la question posée au début du XXème siècle par Willhem Johannsen, en référence à la notion de *force héréditaire* proposée par le semencier français Vilmorin : *On doit décider si la "force héréditaire"* [1]*, comme Vilmorin la nomme, est grande ou petite - cité par Laloë (2011)*. L'évaluation génétique est au coeur des dispositifs d'amélioration génétique ; son but, essentiellement économique, est la sélection d'individus dont l'utilisation comme reproducteur fournira une plus-value économique. Aujourd'hui, cette force héréditaire s'appelle *Expected Progeny Difference*, expression dont les trois termes comptent :

- *Expected* renvoie à la prédiction statistique, c'est-à-dire l'estimation de l'espérance d'une variable aléatoire ;
- *Progeny* renvoie à la transmission génétique entre parents et descendants, et donc à la notion de ressemblance entre apparentés ;
- *Difference* rappelle que l'on cherche à distinguer des individus dans une population.

Une population, pour un généticien, est par essence composée d'individus différents, reliés entre eux par des relations de parenté. Johannsen établit le premier cette distinction essentielle (cité par Laloë (2011)) : *"La théorie statistique de l'hérédité, telle qu'elle est développée par Galton et Pearson* [2]*, s'intéresse aux groupes d'individus ou aux populations.[...] Une population peut contenir de nombreux types indépendants, différant sensiblement les uns des autres, que l'observation de tables ou de courbes de fréquences empirique ne peut absolument pas déceler"*. C'est encore Johannsen, un peu oublié aujourd'hui, qui a forgé les termes *gène*, *phé-*

1. En français dans le texte original allemand
2. Au contraire de Pearson, qui, même s'il est malvenu de le critiquer dans un mémoire consacré en grande part aux analyses factorielles, en était resté à des études de relation entre populations de parents et de descendants.

notype et *génotype*, et a décomposé les variations d'un phénotype en variations transmissibles du génotype et celles non tranmissibles de l'environnement. Fisher (1918) a formalisé cette approche avec son fameux modèle polygénique qui décrit les performances d'une population d'individus par une distribution multinormale dont la matrice de variance-covariance est proportionnelle à la matrice de parenté entre individus [3]. Pour ce faire, il reprend une vieille idée de Mendel selon laquelle une performance est la somme d'un grand nombre d'effets génétiques et suit, par application du théorème central limite, une distribution gaussienne. Les apports d'Henderson actualisent l'approche de Fisher, en lui donnant une traduction achevée sur les plans statistique (modèle mixte et BLUP - *Best Linear Unbiased Prediction* (Henderson, 1975)) et calculatoire (calcul de l'inverse de la matrice de parenté (Henderson, 1976)). Couplés au progrès de l'informatique, ces développements permettent de conduire des évaluations génétiques sur des millions d'animaux, en tenant compte de l'ensemble des relations de parenté entre animaux. Ces grandes populations sont structurées dans l'espace et le temps : les individus naissent à des périodes différentes, dans des endroits différents, et il est difficile de considérer l'environnement comme une simple résiduelle identiquement et indépendamment distribuée. On le modélise généralement par une somme de facteurs à effets fixes. Le modèle classique utilisé en évaluation génétique est donc un modèle mixte gaussien, où une performance est une somme de facteurs environnemnentaux à effets fixes, et d'un facteur génétique à effets aléatoires.

1.1.2 Les indices de sélection

Historiquement, à la mise en place des premières évaluations génétiques, au milieu du XXème siècle, compte tenu de l'état de l'art statistique et informatique, appréhender une population d'individus dans sa caractéristique multidimensionnelle était illusoire. On a donc soit réduit une population à un échantillon indépendamment et identiquement distribué (sélection massale), soit structuré la population en familles d'individus apparentés, le plus souvent des demi-frères (sélection sur descendance), en négligeant toutes les autres relations de parenté. C'est la théorie des indices de sélection. La simplicité de la méthode et sa frugalité informatique pallient largement ses inconvénients, qui sont doubles :

- On ne considère généralement qu'un seul type d'apparentement entre individus, les plus communs étant : individu et lui-même, individu et descendants, individu et collatéraux ;
- la structuration spatio-temporelle de l'environnement est ignorée, ce qui interdit, *sensu stricto*, la comparaison de l'ensemble des individus entre eux.

1.1.3 Le coefficient de détermination - CD

Dans ce contexte, d'où l'aspect multidimensionnel est absent, seule compte la précision individuelle des prédictions, appréhendée par le coefficient de détermination (CD), carré de la corrélation entre valeur prédite $\hat{g}$ et valeur vraie g : $CD = cor^2(g, \hat{g})$. Le CD mesure

3. Pour en finir avec les considérations sémantiques, c'est dans cet article qu'apparaît pour la première fois le mot *Variance*

également un gain d'information, c'est-à-dire une réduction de variance due à l'évaluation : $CD = \frac{var(g|\hat{g})}{var(g)}$. Dans le cas ci-dessus, $CD = h^2$. Enfin, la réponse à la sélection, qui est la supériorité génétique des *p* animaux retenus sur l'ensemble de la population est une fonction croissante simple du CD :

$$R_p = i_p\sqrt{CD}\sigma_g^2 \tag{1.1}$$

où l'intensité de sélection i_p est la différence entre moyenne des prédictions des individus retenus et moyenne de la population. Le CD s'interprète donc doublement :

- statistique : précision de la prédiction, gain d'information
- génétique : efficacité de la sélection

Si l'on considère l'évaluation génétique comme un simple outil, préalable au véritable enjeu qui est le choix des individus qui vont procréer la génération suivante, c'est cette dernière interprétation qui est la plus importante.

1.2 Le modèle mixte et le BLUP

Les inconvénients des indices de sélection disparaissent avec les travaux d'Henderson (1975) sur le modèle mixte, où l'environnement se décompose en une somme de facteurs à effets fixes et d'une résiduelle, et où le vecteur des valeurs génétiques des animaux, aléatoire, suit une loi gaussienne multidimensionnelle. La résolution des équations dites du modèle mixte fournit simultanément les BLUE (*Best Linear Unbiased Estimator*) des facteurs fixes d'environnement, et les BLUP (*Best Linear Unbiased Predictor*) des valeurs génétiques aléatoires. Ces équations font intervenir la matrice inverse des corrélations entre apparentés, pour laquelle Henderson (1976) développe une méthode de calcul très simple, pourvu que l'ensemble des généalogies de la population soit pris en compte. Ces apports ne bouleversent pas les concepts de base évoqués plus haut, mais ils intègrent pleinement l'aspect multidimensionnel dans l'évaluation génétique, et révolutionnent sa pratique. Des algorithmes efficients couplés avec la montée en puissance contemporaine de l'informatique permettent de réaliser des évaluations à des échelles jusque-là inimaginables, de l'ordre du million d'animaux. Ils suscitent également de nouvelles problématiques liées à la structure des plans d'expérience et à la précision générale d'une évaluation. L'estimation simultanée de tous les effets, environnementaux et génétiques, permet la comparaison d'individus issus de différents milieux, ou nés à des périodes différentes. La question de la précision de ces comparaisons, liée en particulier à la disconnexion du plan d'expérience, devient alors pertinente.

Les premières méthodes d'évaluation de la disconnexion dans les évaluations génétiques ont été proposées par Foulley et al. (1990) et Kennedy and Trus (1993). Elles s'appuient toutes deux sur la variance d'erreur de prédiction (PEV). La première introduisait le concept de "degré de connexion", augmentation relative due à la présence d'effets fixes dans le modèle. La seconde suggérait l'étude de la PEV des différences entre candidats à la sélection comme la mesure la plus appropriée. Il s'est trouvé malencontreusement, lors de mes applications sur données réelles, que les résultats fondés sur la PEV quand on prenait la parenté entre indivi-

dus en compte aboutissait à des résultats totalement ininterprétables, et injustifiables auprès des utilisateurs finaux. C'est pour cela que j'ai développé un concept généralisant le coefficient de détermination à un contexte multidimensionnel.

1.2.1 Eléments théoriques.

On considère un modèle mixte gaussien avec un facteur à effets aléatoires

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \tag{1.2}$$

où **y** est le vecteur des performances, **b** le vecteur des effets fixes, **u** le vecteur des effets aléatoires de dimension n, **e** le vecteur des résiduelles, **X** et **Z** sont les matrices d'incidence. (**y,u,e**)' suit la distribution suivante :

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{e} \end{pmatrix} \sim N \left[\begin{pmatrix} \mathbf{Xb} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{ZAZ'}\sigma_a^2 + \mathbf{I}\sigma_e^2 & \mathbf{ZA}\sigma_a^2 & \mathbf{I}\sigma_e^2 \\ \mathbf{AZ'}\sigma_a^2 & \mathbf{A}\sigma_a^2 & \mathbf{0} \\ \mathbf{I}\sigma_e^2 & \mathbf{0} & \mathbf{I}\sigma_e^2 \end{pmatrix} \right] \tag{1.3}$$

où **A** est la matrice de parenté entre les animaux, σ_a^2 et σ_e^2 sont les variances génétique et résiduelle. On note $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ le projecteur orthogonal au sous-espace vectoriel engendré par les colonnes de **X**, (*ie* **MX=0**). Le BLUP des effets **u**, noté $\hat{\mathbf{u}}$ est la solution de l'équation (Henderson, 1975) :

$(\mathbf{Z}'\mathbf{M}\mathbf{Z}+\lambda\mathbf{A}^{-1})\hat{\mathbf{u}} = \mathbf{Z}'\mathbf{M}\mathbf{y}$

La distribution conjointe de **u** et $\hat{\mathbf{u}}$ est multinormale, d'espérance nulle et de variance :

$$Var \begin{pmatrix} \mathbf{u} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \Psi \\ \Psi & \Psi \end{pmatrix} \sigma_a^2 \tag{1.4}$$

avec $\Psi = \mathbf{A} - \lambda(\mathbf{Z}'\mathbf{M}\mathbf{Z}+\lambda\mathbf{A}^{-1})^{-1}$, et $\lambda = \dfrac{\sigma_e^2}{\sigma_a^2}$. Enfin, la variance d'erreur de prédiction (PEV), critère classique pour juger de la précision d'une prédiction, s'écrit :
$PEV = var(\mathbf{u} - \hat{\mathbf{u}}) = (\mathbf{Z}'\mathbf{M}\mathbf{Z}+\lambda\mathbf{A}^{-1})^{-1}\sigma_e^2$. Ce modèle se décline en plusieurs variantes selon la composition de **u**. Les plus courantes sont le modèle individuel, où les performances sont reliées à la valeur génétique de l'individu et le modèle père, où les performances sont reliées à la valeur génétique du père.

1.2.2 Généralisation du CD.

Sauf mention contraire, tous les résultats présentés dans cette section et les suivantes sont tirés de (Laloë (1993), Laloë et al. (1996), Laloë and Phocas (2003).

Dans le cadre unidimensionnel d'un indice de sélection, la précision de l'évaluation d'un animal s'appréhende par le CD, carré de la corrélation entre valeur prédite et valeur vraie, et

encore la réduction de variance due à l'évaluation. La généralisation du CD au cadre multidimensionnel passe, si l'on retient l'interprétation en terme de carré de corrélation, par l'analyse des corrélations canoniques, qui compare deux ensembles de variables mesurées sur les mêmes individus, et consiste à diagonaliser $\mathbf{A}^{-1}\Psi\Psi^{-1}\Psi = \mathbf{A}^{-1}\Psi$. La deuxième interprétation revient à comparer deux formes quadratiques $CD(\mathbf{x}) = \dfrac{\mathbf{x}'\mathbf{\Psi}\mathbf{x}}{\mathbf{x}'\mathbf{A}\mathbf{x}}$. Dans les deux cas, on résout l'équation généralisée aux valeurs propres :

$$[\mathbf{\Psi} - \mu\mathbf{A}]\,\boldsymbol{\beta} = \mathbf{0} \tag{1.5}$$

Les vecteurs propres $\boldsymbol{\beta}_k$ et les valeurs propres μ_k, triées par ordre croissant, sont telles que, pour $i \neq j$:

$$\boldsymbol{\beta}_i'\mathbf{A}\boldsymbol{\beta}_j = 0 \tag{1.6}$$
$$\boldsymbol{\beta}_i'\mathbf{A}\boldsymbol{\beta}_i = 1 \tag{1.7}$$
$$\boldsymbol{\beta}_i'\mathbf{\Psi}\boldsymbol{\beta}_j = 0 \tag{1.8}$$
$$\boldsymbol{\beta}_i'\mathbf{\Psi}\boldsymbol{\beta}_i = \mu_i \tag{1.9}$$
$$\forall \mathbf{x} \neq \mathbf{0}, \mu_1 \leq CD(\mathbf{x}) \leq \mu_n \tag{1.10}$$

A étant définie positive [4], **L** matrice triangulaire inférieure non singulière existe telle que **A=LL'**. On a alors :

$$(1.5) \Leftrightarrow \left[\mathbf{L^{-1}\Psi} - \mu\mathbf{L^{-1}A}\right]\boldsymbol{\beta} = \mathbf{0} \tag{1.11}$$
$$\Leftrightarrow \mathbf{L^{-1}\Psi}\boldsymbol{\beta} = \mu\mathbf{L^{-1}A}\boldsymbol{\beta} \tag{1.12}$$
$$\Leftrightarrow \mathbf{L^{-1}\Psi}\boldsymbol{\beta} = \mu\mathbf{L^{-1}LL'}\boldsymbol{\beta} \tag{1.13}$$
$$\Leftrightarrow \mathbf{L^{-1}\Psi L'^{-1}c} = \mu\mathbf{c} \tag{1.14}$$

D'où :

$$[\mathbf{\Theta} - \mu\mathbf{I}]\,\mathbf{c} = \mathbf{0} \tag{1.15}$$

avec $\mathbf{c} = \mathbf{L}'\boldsymbol{\beta}$ et $\mathbf{\Theta} = \mathbf{L^{-1}\Psi L'^{-1}}$.

Les équations (1.5) et (1.15) ont les mêmes valeurs propres. D'autre part, $\mathbf{\Theta}$ peut s'écrire sous la forme $\mathbf{\Theta} = \mathbf{I} - (\lambda^{-1}\mathbf{L}'(\mathbf{Z}'\mathbf{M}\mathbf{Z})\mathbf{L} + \mathbf{I})^{-1}$. On en déduit que

- $\mathbf{\Theta}$ et **L'Z'MZL** ont les mêmes vecteurs propres car $\mathbf{\Theta}$ est une fonction linéaire de **I** et de l'inverse d'une fonction linéaire de **I** et **L'Z'MZL**.
- Les CD sont compris entre 0 et 1, car, si pour un vecteur propre donné, la valeur propre correspondante de **L'Z'MZL** est ν, alors la valeur propre correspondante μ de $\mathbf{\Theta}$ est égale à $\mu = 1 - \dfrac{1}{1+\lambda^{-1}\nu}$, ou encore à :

$$\mu = \frac{\lambda^{-1}\nu}{1+\lambda^{-1}\nu} \tag{1.16}$$

Z'MZ étant non négative et **A** positive, **Z'MZA** et **L'Z'MZL** sont non négatives. Par suite, $\nu \geq 0$ et donc : $0 \leq \mu < 1$.

4. Si l'on exclut le cas des clones et des jumeaux

- $\boldsymbol{\Theta}$ et **Z'MZ** ont le même rang, car $\boldsymbol{\Theta}$ et **L'Z'MZL** ont les mêmes vecteurs propres, et de (1.16), on déduit qu'à une valeur propre nulle de $\boldsymbol{\Theta}$ correspond une valeur propre nulle de **L'Z'MZL**. Ces deux matrices ont donc le même rang. Enfin, **L** étant non singulière, **L'Z'MZL** et **Z'MZ** ont le même rang *r*.
- $r \leq n - 1$. On a en effet **Z'MZ1=Z'M1**, car **Z1** est le vecteur des sommes des lignes de **Z**, et donc égal à **1**. D'autre part, **M1=0**, puisque **1** est combinaison linéaire des colonnes de **X**, et que **M** est le projecteur orthogonal à l'espace vectoriel engendré par ces colonnes. Donc **Z'MZ** admet toujours une valeur propre nulle, associée au vecteur propre $\mathbf{1}/\sqrt{n}$. Les autres vecteurs propres sont des contrastes orthogonaux à **1**.
- De façon similaire, l'équation (1.5) admet toujours une valeur propre $\mu_0 = 0$, correspondant à un vecteur propre $\mathbf{c}_0$ proportionnel à $\mathbf{A^{-1}1}$. Les autres vecteurs propres $\{\mathbf{c}_i\}_{i=1,n-1}$ sont $\mathbf{A}$-orthogonaux à $\mathbf{A^{-1}1}$, et sont donc des contrastes.
- On se restreindra donc dans la suite de l'étude au sous-espace vectoriel des contrastes.

1.2.3 Critères généraux de précision.

On déduit de l'intervalle (1.10) des CD des critères globaux de précision comme les moyennes arithmétique et géométrique des *(n-1)* plus grandes valeurs propres de $\boldsymbol{\Theta}$:

$$\rho_1 = \frac{tr(\Theta)}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n-1}\mu_i \tag{1.17}$$

$$\rho_2 = \left[\prod_{i=2}^{n-1}\mu_i\right]^{\frac{1}{n-1}} \tag{1.18}$$

On notera l'analogie entre ces deux critères et les critères intervenant dans l'optimisation des plans d'expérience, fondés sur la trace (A-optimalité) et le déterminant (D-optimalité) de la matrice d'information associée au plan d'expérience.

1.2.4 Information fournie par l'évaluation génétique.

Une approche complémentaire utilise le concept d'information fournie par l'évaluation, par exemple au moyen de l'information de Kullback, qui mesure la distance entre deux distributions. L'information qui nous intéresse est celle apportée en moyenne par l'évaluation, définie par le plan d'expérience (**X** et **Z**). Pour un contraste **x**, l'information $I_{\mathbf{x}}$ est fonction du CD $CD(\mathbf{x})$:

$CD(\mathbf{x}) = 1 - \exp(-2I_{\mathbf{x}})$

L'information est une fonction croissante du CD, nulle quand le CD est nul, et tendant vers l'infini quand le CD tend vers 1. A partir de ce résultat,on montre qu'un contraste à CD nul, et donc à information nulle, est toujours nul, quelles que soient les performances : l'évaluation génétique n'apporte aucune information sur ce contraste, et "fixe" sa valeur à 0 :

$$\forall \mathbf{y}, CD(\mathbf{x}) = 0 \Leftrightarrow \mathbf{v}'\mathbf{A^{-1}}\hat{\mathbf{u}} = 0 \tag{1.19}$$

1.2.5 Ecriture des contrastes.

En résumé, l'étude de la précision d'un dispositif passe par l'étude de la précision des contrastes entre effets. Ces contrastes sont dans un sous-espace vectoriel, de dimension *n-1* et dont une base $(\mathbf{c}_i)_{i=1,n-1}$ est formée des contrastes vecteurs propres de (1.5). Nous apppellerons ces contrastes "contrastes canoniques", et leurs CD, "CD canoniques", en référence à l'analyse des corrélations canoniques. Tout contraste $\mathbf{x}$ est combinaison linéaire des constrastes canoniques :

$$\mathbf{x} = \sum_{i=1}^{n-1} a_i \mathbf{c}_i \tag{1.20}$$

et, du fait des propriétés d'orthogonalité liées à la décomposition canonique, son CD est une moyenne pondérée des CD canoniques :

$$CD(\mathbf{x}) = \frac{\sum_{i=1}^{n-1} a_i^2 CD_i}{\sum_{i=1}^{n-1} a_i^2} \tag{1.21}$$

1.2.6 La disconnexion.

Fondamentalement, la disconnexion caractérise la structure d'un plan d'expérience. Elle se produit quand des facteurs sont confondus ou emboîtés. Elle se visualise simplement dans le cas d'un plan à deux facteurs (1.1) : un plan est disconnecté si on ne peut pas relier toutes les cellules contiguës avec observation (symbolisé par un "X") par un trait horizontal ou vertical. Cette caractérisation n'est plus possible dans les modèles plus compliqués, et on préfère définir la disconnexion par ses conséquences en terme de rang de matrices ou de contrastes inestimables.

Elevage Taureau	V1	V2	V3
T1	X		
T2	X		
T3		X	
T4		X	
T5			X
T6			X

TABLE 1.1 – Un plan d'expérience disconnecté

1.2.7 La disconnexion dans un modèle à effets fixes.

La disconnexion a d'abord été définie pour les modèles à effets fixes. Dans un plan disconnecté, certains contrastes ne sont plus estimables. L'ensemble des contrastes estimables est

l'espace vectoriel dont une base est constituée des vecteurs propres de **Z'MZ** associés aux valeurs propres non nulles de cette matrice. La dimension de l'espace des contrastes estimables, égale au rang de **Z'MZ**, est inférieure ou égale à *n-1*. Un système de contraintes d'identifiabilité rendant le modèle de plein rang est alors l'annulation des combinaisons linéaires de $\mathbf{u}$ associées aux vecteurs propres nuls de **Z'MZ** (Coursol, 1980). En clair, on force à 0 les contrastes non estimables. Si $\mathbf{B^0}$ est la matrice constituée des vecteurs propres de **Z'MZ** associés aux valeurs propres nulles, ce système de contraintes s'écrit $\mathbf{B^0 u} = \mathbf{0}$.

1.2.8 La disconnexion des effets aléatoires dans un modèle mixte.

Une caractérisation semblable pour les effets aléatoires s'obtient par l'examen de l'incidence de la structure des données sur les valeurs et vecteurs propres de (1.15). Quand ils sont de valeur propre nulle, les vecteurs propres de (1.5) se déduisent des vecteurs propres de **Z'MZ** (Foulley et al., 1990) :

$$\mathbf{Z'MZb} = \mathbf{0} \Leftrightarrow \mathbf{\Psi A^{-1} b} = \mathbf{0} \Leftrightarrow CD(\mathbf{b}) = 0$$

$\mathbf{b'A^{-1}\hat{u}}$ est donc toujours nul, d'après (1.19), et ce pour tout $\mathbf{y}$. Ces équations forment un système analogue au système de contraintes d'identifiabilité évoqué ci-dessus et établi pour rendre de plein rang un modèle à effets fixes. Dans le cas des effets aléatoires, ce système d'annulation de contrastes est implicite.
Dans un modèle mixte, la disconnexion entre effets aléatoires se traduit donc par l'existence de contrastes à CD nul.

1.2.9 La géométrie des contrastes.

A partir de ces équations, on partitionne l'espace vectoriel des contrastes en somme de deux sous-espaces vectoriels, l'un généré par la famille des vecteurs propres de valeur propre nulle (contrastes non estimables, ou à CD nul), et l'autre par la famille des autres vecteurs propres, de valeur propre positive (contrastes estimables, ou à CD positifs). La situation est par contre différente pour les contrastes n'appartenant pas à ces deux sous-espaces (Figure 1.1). Ils sont inestimables dans le cas fixe, alors que leur CD est positif pour les effets aléatoires, puisqu'ils sont une moyenne pondérée de CD nuls et de CD positifs (cf (1.21)).
Cela peut s'expliquer si on raisonne en terme d'apport d'information par rapport à une connaissance préalable. On ne dispose d'aucune information préalable sur la distribution d'un effet fixe, alors que, pour un effet aléatoire, cette information se modélise simplement par une loi gaussienne. En interprétant la disconnexion comme un apport nul d'information pour un contraste, les distributions a posteriori ne sont pas modifiées. Dans le cas aléatoire, cela se traduira simplement, dans (1.21), par de nombreuses valeurs annulées, donc une régression vers la moyenne, et une baisse de CD, d'autant plus importante que le poids de contrastes à CD nuls sera fort. Par contre, dans le cas fixe, cela aboutit à une impossibilité de calcul.

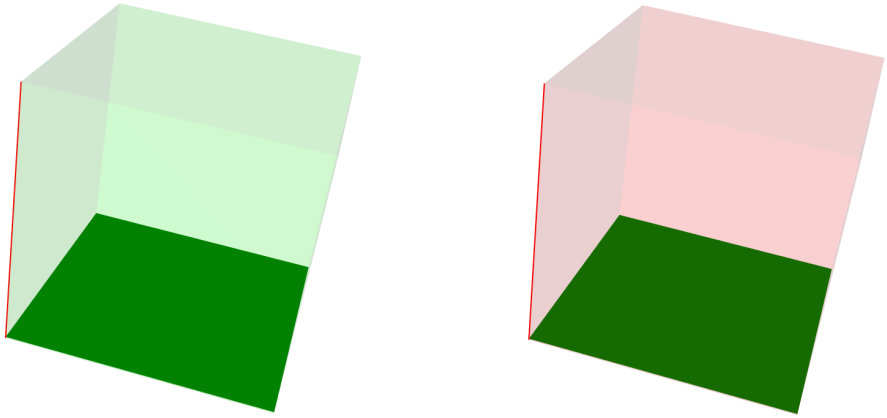


FIGURE 1.1 – Géométrie des contrastes. A gauche, facteurs à effets aléatoires ; à droite, facteurs à effets fixes. Le cube matérialise le sous-espace vectoriel des contrastes. La droite rouge matérialise le sous-espace vectoriel engendré par la famille des vecteurs propres de valeur propre nulle, et le plan vert son supplémentaire, engendré par la famille des vecteurs propres de valeurs propres positives. L'interprétation de ces deux sous-espaces vectoriels est similaire : contrastes estimables, ou de CD positif d'une part, contraste non-estimable ou de CD nul d'autre part. Pour les contrastes n'appartenant à aucun de ces deux sous-espaces, la situation est par contre différente. Ils sont non-estimables dans le cas fixe, alors que leur CD est positif dans le cas aléatoire.

1.2.10 Etude de dispositifs.

L'étude de dispositifs simples, pour lesquels l'équation (1.5) se résout analytiquement, permet de préciser l'impact de la parenté ou de la structure d'un dispositif sur l'efficacité de la sélection.

1.2.10.1 Animaux autocorrélés.

Pour une population d'individus tous liés entre eux par le même coefficient de parenté *r*, tous les CD de contrastes sont égaux à $\frac{1-r}{1-r+\lambda}$; la réponse à la sélection s'écrit $R_{p,r} = i_{p,r}\sqrt{CD}\sigma^2$, où $i_{p,r} = i_p\sqrt{(1-r)}$ est l'intensité de sélection tenant compte du coefficient de parenté *r* entre individus. On notera la similitude avec (1.1). L'apparentement diminue la réponse à la sélection, et donc son efficacité. La prédiction des différences est certes plus précise (le PEV baisse avec la parenté), mais la variabilité génétique, et donc les différences entre individus, sont plus faibles : on prédit plus précisément des différences plus faibles. Au contraire, le CD est proportionnel à la réponse à la sélection.

1.2.10.2 Dispositif hiérarchique.

Un autre cas est, pour un modèle père, un dispositif hiérarchique décrit en figure 1.2, avec *N* "cellules", généralement des troupeaux ou des combinaisons "troupeau*période", des pères n'ayant des descendants que dans une cellule (*s* pères par cellule, et n_p descendants /père*cellule), et *t* pères "connecteurs", avec des descendants dans toutes les cellules (n_q descendants /père*cellule)), avec $\nu = \dfrac{tn_q}{sn_p + tn_q}$ notant la proportion de descendants des pères connecteurs. En se restreignant aux pères intra-cellules, on obtient deux types de contrastes

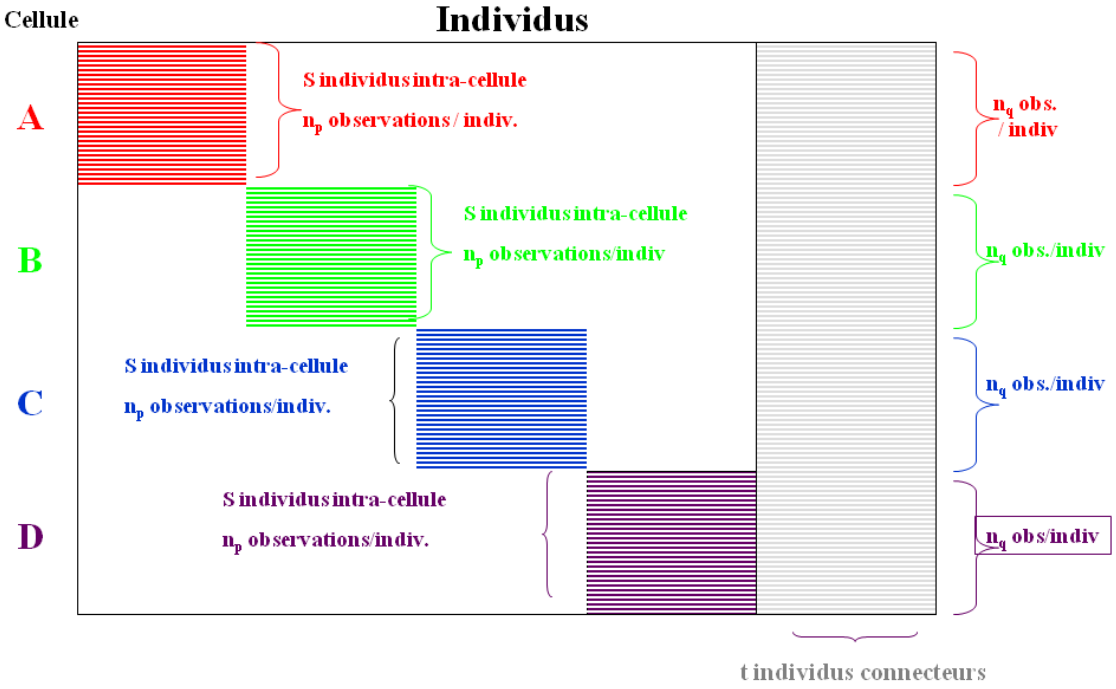


FIGURE 1.2 – Dispositif hiérarchique.

canoniques, les contrastes intra-cellule et les contrastes entre cellules. Leurs CDs sont indiqués dans la table 1.2. La formule $CD_b = \dfrac{n_p\nu}{n_p\nu + \lambda}$ synthétise les trois facteurs intervenant dans la précision, à savoir la quantité d'information associée à un individu (n_p), la qualité du dispositif (ν), et l'héritabilité, fonction de λ. Enfin, elle objective le lien entre connexion, mesurée par ν, et la précision des comparaisons entre individus.

Type de contraste	Nbre de contrastes	Contraste canonique	CD
intra-cellule	$N(s-1)$	$\sqrt{\frac{s}{s-1}}[u_{A_i} - \bar{u}_A]$	$CD_w = \frac{n_p}{n_p + \lambda}$
Entre cellules différentes	$N-1$	$\sqrt{\frac{s}{2}}[\bar{u}_A - \bar{u}_B]$	$CD_b = \frac{n_p \nu}{n_p \nu + \lambda}$

TABLE 1.2 – CD canoniques

1.2.11 Le biais dans les évaluations génétiques.

Une hypothèse sous-jacente à l'utilisation du BLUP en évaluation génétique est, qu'en espérance, les valeurs génétiques des animaux sont égales. C'est une hypothèse lourde en cas de structuration importante spatio-temporelle. Le but de la sélection génétique est d'augmenter le niveau génétique au cours du temps, et on peut difficilement admettre l'égalité en espérance des valeurs génétiques d'individus nés à des périodes différentes ; de même dans les évaluations internationales, où l'on compare des individus de pays différents, et où on peut supposer une différence systématique entre niveaux génétiques de pays différents. Ces différences systématiques sont à l'origine de biais dans les prédictions des contrastes entre individus de pays ou de périodes différents. Si l'étude de la précision passe par la distribution de $\mathbf{u}|\hat{\mathbf{u}}$, l'étude du biais passe par la distribution de $\hat{\mathbf{u}}|\mathbf{u}$, qu'on peut interpréter en terme de puissance : connaissant les différences "vraies" entre individus, que peut-on dire des différences prédites. On retrouve, en reprenant les notations de (1.20) l'expression de la régression vers la moyenne, d'autant plus sévère que le CD est faible :

$$E(\mathbf{x}'\hat{\mathbf{u}}|\mathbf{x}'\mathbf{u}) = \sum_{i=1}^{n-1} a_i CD_i \mathbf{c}_\mathbf{i}'\mathbf{u} \tag{1.22}$$

D'où la question : s'il existe une différence a priori entre niveaux génétiques de deux pays, comment cette différence est-elle prédite dans l'évaluation. Reprenons le dispositif hiérarchique qui, même s'il peut paraître simpliste, reproduit assez fidèlement un dispositif d'évaluation génétique, avec une majorité de pères utilisés dans un seul pays, et des pères utilisés dans des pays différents assurant la connexion. La table 1.3 donne les expressions du CD et du biais du contraste entre deux pères de deux cellules différentes. Ce contraste est somme de trois contrastes canoniques, les deux contrastes intra-cellule entre individu et moyenne génétique de la cellule, et le contraste entre les niveaux génétiques des cellules (en rouge), ce dernier pouvant être considéré comme d'espérance non nulle. Une connexion nulle ou faible entre cellules a deux conséquences sur un contraste entre deux pères de cellules différentes :

- La précision décroît, mais de façon négligeable dès que le nombre de pères dépasse quelques dizaines ;
- un biais est probable si les niveaux génétiques des cellules diffèrent, et ce, quelle que soit la taille des cellules.

En conclusion, un contraste à CD élevé peut tout à fait être biaisé. L'étude du biais potentiel d'un contraste passe par son écriture en fonction des contrastes canoniques.

Comparaison	Décomposition
$[u_{A_i} - u_{B_j}]$	$[u_{A_i} - \bar{u}_A] - [u_{B_j} - \bar{u}_B] + [\bar{u}_A - \bar{u}_B]$
CD	$\frac{s-1}{s} CD_w + \frac{1}{s} CD_b$
Biais	$(1 - CD_w) \underbrace{\{E[u_{A_i} - \bar{u}_A] - E[u_{B_j} - \bar{u}_B]\}}_{=0} + (1 - CD_b) \underbrace{E[\bar{u}_A - \bar{u}_B]}_{\neq 0}$

TABLE 1.3 – Comparaison de deux individus de cellules différentes

1.3 Applications et développements.

Schématiquement, on peut distinguer deux types d'évaluation génétique selon le contexte de recueil des performances des individus : le contrôle en station, et le contrôle en ferme.

1.3.1 Etude des plans d'expérience dans les stations de contrôle des bovins allaitants

Dans le premier cas, les données proviennent de stations dites de contrôle. Elles concernent un faible nombre d'individus (de l'ordre de la centaine), et sont recueillies selon un protocole planifié, précis et maîtrisé. On est ici dans la problématique classique de l'optimisation d'un plan d'expérience : on choisira tel ou tel type de dispositif en fonction d'un critère d'optimalité, souvent relié à une caractéristique de la matrice de variance-covariance des estimations. Nous avons ainsi étudié, avec l'aide d'un étudiant de Master (Dodelin et al., 2000), deux plans d'expérience utilisés dans des stations de contrôle où des jeunes taureaux sont évalués à partir des performances de leurs descendants. A l'issue des évaluations annuelles, les meilleurs taureaux sont agréés à l'insémination artificielle, et rejoignent un pool de taureaux évalués les années précédentes. Si, ce qui est souhaitable, un progrès génétique existe, les valeurs génétiques des taureaux dépendent de leur âge et de la série dans laquelle ils ont été évalués : en espérance, les jeunes taureaux sont meilleurs que les vieux. Quel dispositif de connexion entre séries permet de réduire le biais dans la prédiction des différences entre taureaux évalués dans des séries différentes ? Deux dispositifs de connexion sont comparés, un dispositif "taureau de référence" où un taureau dit de référence relie toutes les séries et un dispositif "demi série" où certains taureaux sont contrôlés pendant deux années consécutives (cf Figure 1.3). L'approche décrite dans le paragraphe 1.2.11, confirmée par des résultats de simulation, permet d'établir une relation directe entre CD entre niveaux génétiques des séries et prédiction du progrès génétique. On en conclut à la supériorité du dispositif "taureau de référence".

TAUREAU / ANNEE	T1	T2	T3	T4	T5	T6	T. réf
Année 1	x	x					x
Année 2			x	x			x
Année 3					x	x	x

TAUREAU / ANNEE	T1	T2	T3	T4	T5
Année 1	x	x	x		
Année 2		x	x	x	
Année 3			x	x	x

FIGURE 1.3 – Les deux dispositifs d'évaluation des taureaux. A gauche, le dispositif "taureau de référence ; à droite, le dispositif "demi-série"

1.3.2 Identification des animaux connectés dans une évaluation génétique : Le CACO

Dans le contrôle en ferme, les données sont recueillies sur le terrain, dans des exploitations agricoles. Elles sont en très grand nombre (jusqu'à plusieurs millions), très hétérogènes quant à leur qualité, le lieu et la période de leur recueil. Il est ici hors de question d'optimiser une structure de données sur laquelle on a une maîtrise réduite, et où la taille du dispositif interdit le calcul de critères d'optimalité à partir d'une analyse des corrélations canoniques. Il s'agira de s'appuyer sur un cadre théorique pour préconiser des stratégies d'amélioration du dispositif, de concevoir des approches permettant de distinguer dans la masse d'individus évalués, ceux suffisamment précis et comparables. Nous avons proposé, dans la suite du doctorat de Marie-Noëlle Fouilloux (Fouilloux et al., 2008a), un critère approché, le *CACO*, pour *Critère d'Admission au rang des troupeaux COnnectés*. C'est une méthode en deux temps :

- On estime d'abord, par simulation, les matrices de variance-covariance empiriques des valeurs génétiques vraies et prédites, selon une méthode proposée par Fouilloux and Laloë (2001). Concrètement ne sont calculés que les CD des contrastes entre niveaux génétiques des cellules.
- On construit ensuite des clusters de cellules pour lesquelles ces CD sont tous supérieurs à un seuil donné. C'est une démarche analogue à la méthode du lien complet, mais adaptée pour le traitement de grands effectifs.

Cette méthode est utilisée en France, en Espagne, au Brésil et en Norvège (Fouilloux et al. (2008a), Tarres et al. (2010), Pegolo et al. (2012), Andonov et al. (2014))

1.3.3 Le biais dans les évaluations internationales

La mondialisation n'ayant pas épargné les programmes de sélection, l'échelle des évaluations peut être encore plus grande. Le rayonnement de l'insémination artificielle, les progrès du transfert embryonnaire et le développement des méthodes de cryoconservation ont stimulé les échanges internationaux de matériel génétique (animaux, semence, embryon), provoquant du même coup un besoin de comparer les reproducteurs à un niveau supranational (Leclerc, 2008). Ce besoin s'est d'abord manifesté chez les bovins laitiers, aboutissant à la création d'Interbull, organisme chargé entre autres de l'évaluation internationale des bovins laitiers, et qui a suscité par la suite plusieurs émules chez les bovins allaitants (Interbeef) ou les chevaux (Interstallion). L'importance des enjeux, notamment économiques, impose à Interbull d'assurer des comparaisons équitables entre reproducteurs de tous les pays. Cela repose, entre autres,

sur des biais de prédiction limités entre animaux de pays différents. Si l'on se réfère à la discussion précédente sur les biais (paragraphe 1.2.11) dans un dispositif hiérarchique équilibré (paragraphe 1.2.10.2), nous sommes dans un cas où les effectifs intra-cellule (le pays) sont grands, et où la détermination des contrastes canoniques est impossible. Le calcul des CD ne permettra pas de déduire quoi que ce soit sur les risques de biais. Le plus pertinent sera donc la recherche directe de biais potentiels. Toujours avec Marie-Noëlle Fouilloux (Fouilloux et al., 2008b), nous avons développé une méthode fondée sur des simulations intégrant des différences systématiques de niveau génétique entre pays, et où le risque de biais est quantifié par le rapport des différences prédites aux différences vraies.

1.3.4 Un second souffle avec la sélection génomique ? Le choix du set d'apprentissage

L'irruption de la sélection génomique au début des années 2000 (Meuwissen et al., 2001) et une réorientation de mes activités au sein de GABI m'avait convaincu de me détourner de cette thématique. Mais il se trouve que ces recherches ont rencontré un écho dans la communauté de la génétique végétale...
En sélection génomique, les valeurs génétiques ne sont plus prédites à partir de performances et de relations de parenté, mais à partir des génotypes établis sur un grand nombre de marqueurs moléculaires. C'est la concrétisation du modèle polygénique de Fisher. Une formule de prédiction est développée à partir des génotypes et phénotypes d'un ensemble d'individus (*calibration set* ou *training set - échantillon d'apprentissage*). La taille et la composition de cet ensemble est un facteur clé, dans la mesure où elle exerce une grande influence sur la précision de la sélection génomique. Dans le cadre du doctorat de Renaud Rincent, une étude à laquelle j'ai contribué, Rincent et al. (2012) compare plusieurs critères, fondés sur le PEV et le CD, en considérant les moyennes des PEV (PEVmean) ou des CD (CDmean) des contrastes entre chaque candidat et la moyenne de la population. Le test de ces critères sur deux panels de maïs, a montré la supériorité du CDmean, en particulier du fait de la prise en compte de la parenté dans la procédure. De même, Isidro et al. (2014) ont comparé plusieurs algorithmes de choix de l'échantillon d'apprentissage fondés sur différents critères, et a montré sur des populations de blé la supériorité du CD. Pour ces auteurs, le fait que ces algorithmes minimisent la parenté intra-set tout en maximisant les relations de parenté entre set d'apprentissage et set de test le rend approprié pour la sélection à long terme. Enfin, au terme d'une étude sur l'huile de palme, Cros et al. (2014) en arrivent à la même conclusion quant à la supériorité des algorithmes fondés sur le CD.

1.3.5 Conclusion

L'approche géométrique utilisant l'analyse des corrélations canoniques permet d'appréhender les différents aspects liés à l'efficacité d'une évaluation génétique (précision, connexion, biais, réponse à la sélection), en respectant :

- les spécificités multidimensionnelle et aléatoire des effets génétiques, et en particulier tous les aspects liés aux liens de parenté entre individus ;
- la finalité de l'évaluation génétique, qui cherche à révéler les différences. Le CD est cohérent avec la réponse à la sélection, en particulier en présence d'apparentements entre individus : l'apparentement diminue la variance d'erreur de prédiction, mais diminue également l'intensité de sélection et la variabilité génétique, ce qui se traduit par la baisse conjointe de la réponse à la sélection et du CD. Cette relation entre réponse à la sélection et CD, mise en évidence sur des cas théoriques, a été confirmée par d'autres études sur des cas réels (e.g. Kuehn et al. (2007)).

Si l'analyse des corrélations canoniques est impossible pour des évaluations génétiques de grande taille, elle a été un support pour le développement de méthodes approchées, que ce soit pour déterminer des ensembles d'individus connectés, ou les biais potentiels dans la prédiction des valeurs génétiques de reproducteurs.

2

Analyse géométrique des données en génétique des populations

Au début de la décennie 1990, quand je commence à travailler sur la thématique de structuration génétique des populations dans le cadre de la thèse de Katayoun Moazami-Goudarzi, la démarche classique était de calculer des distances génétiques et à visualiser des relations entre populations par des arbres, dont la stabilité était appréhendée par bootstrap (Felsenstein, 1985). Cette thèse avait pour objet l'analyse de dix races bovines françaises à l'aide de 17 microsatellites. Les résultats (Moazami-Goudarzi et al., 1997) ont été assez décevants, avec des arbres peu robustes, et sans qu'une structure claire, à savoir des groupes de populations bien individualisés, ne se dégage. De plus, l'augmentation du nombre de marqueurs ne permet aucune amélioration. A mon sens, la démarche "Construction d'arbres + Bootstrap" ne permet pas d'analyser cet état de fait, et ce, pour deux raisons :

- Vu le faible nombre de marqueurs, l'examen des structures induite par chaque microsatellite et de leurs ressemblances est plus efficace qu'une procédure de rééchantillonnage qui, par construction, n'individualise pas les effets des marqueurs.
- L'analyse et la quantification des ressemblances entre structures individuelles n'est pas chose aisée à partir de dendrogrammes.

C'est pourquoi je me suis tourné vers les analyses factorielles. L'équipe de Cavalli-Sforza (Cavalli-Sforza, 1966) avait déjà utilisé l'Analyse en Composantes Principales (ACP), qui présentait à leurs yeux plusieurs avantages, entre autres l'absence d'un modèle de différenciation génétique et la possibilité de représenter des variations graduelles (clines). Pour nous, c'était la possibilité, grâce aux contributions des allèles dans la construction des composantes principales, de quantifier le rôle respectif des microsatellites dans l'élaboration de la structure finale (Moazami-Goudarzi and Laloë, 2002), qui nous a convaincu de l'utiliser. En particulier, un cercle des corrélations (cf Figure 2.1) issu d'une ACP sur les distances entre populations calculées à partir des microsatellites individuels, nous a permis d'apprécier visuellement le niveau de congruence des structures construites par chaque microsatellite : chaque point représente une typologie de populations induite par un marqueur, et la proximité entre deux points indique une similarité entre les typologies correspondantes. Appliquée à des populations de

taurins européens, les microsatellites sont répartis un peu partout à l'intérieur du cercle (Figure 2.1 - (A)), traduisant une incongruence entre les structures individuelles. Espérer dans ce cas un résultat consensuel n'est pas raisonnable. Au contraire, appliquée à des bovins africains et européens, à la structure plus marquée, les microsatellites sont très majoritairement proches les uns des autres, indiquant que ces microsatellites construisaient des structures similaires (Figure 2.1 - (B)).

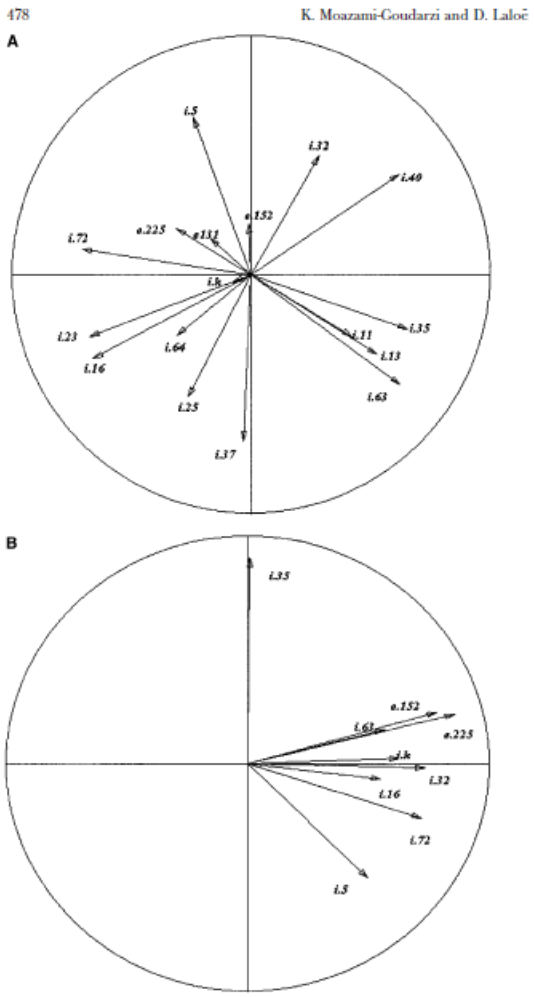


FIGURE 2.1 – Des structures individuelles non congruentes (Moazami-Goudarzi and Laloë, 2002). Des individus de différentes populations sont génotypés sur plusieurs microsatellites. On construit un tableau de distances génétiques pour chaque paire de populations (en ligne), et chaque microsatellite (en colonne). Une ACP est faite sur ce tableau ; un cercle de corrélations visualise les relations entre les microsatellites. Cette procédure est appliquée sur (A) des populations de taurins européens génotypés sur 17 microsatellites ; (B) des populations de deux sous-espèces différentes (taurins *vs* zébus), réparties sur deux continents (Afrique, Europe)

.

Cette première incursion dans l'univers des analyses factorielles a été l'occasion de rencontrer le professeur Chessel, du laboratoire de Biologie et Biométrie Evolutive de l'université de Lyon. A son contact, j'ai découvert le formalisme du schéma de dualité (Escoufier, 1987), et le prolongement des méthodes de base vers des outils plus complexes.
Les analyses factorielles, dans leur variété, permettent de répondre à des questions générales, qui débordent le cadre de la génétique des populations. Ces questions sont de trois ordres :

Structuration des variables. Derrière ce terme un peu vague, se cachent deux grandes interrogations : tout d'abord le problème de la construction d'un consensus et de sa signification, que nous avons déjà rencontré. En second lieu, l'appréhension exhaustive d'un phénomène par une multiplicité de points de vue, l'intégration de données hétérogènes, c'est-à-dire la description synthétique d'un phénomène sous ses différents aspects.

Lien entre diversité et marqueurs. Déjà présent dans notre interrogation sur l'établissement d'un consensus, ce point a gagné de l'importance avec l'apparition de marqueurs génétiques localisés sur le génome, auxquels on peut associer de plus en plus de gènes à l'action connue.

A toutes ces questions, les analyses factorielles apportent des réponses :

- Consensus et *analyse de coinertie multiple*, intégration et *analyse factorielle multiple*, ces deux analyses appartenant aux méthodes k-tableaux. Les k-tableaux sont des structures de données constituées de plusieurs tableaux ayant en commun les lignes ou les colonnes.
- Des ACP ont déjà été utilisées pour interpréter la diversité par la géographie (e.g. Novembre and Di Rienzo (2009)). Ces auteurs utilisent les scores des populations ou des individus sur les premiers axes d'une ACP. Mais cette procédure souffre d'un défaut important : elle s'appuie sur une information synthétique pour tenter de déceler des structures qui ont pu échapper à la synthèse. En clair, l'information résumée sur les premiers axes peut ne pas être structurée spatialement. C'est pourquoi il est préférable d'utiliser des procédures ad hoc, telles que *l'analyse spatiale*, qui intègrent l'information géographique via des graphes de voisinage et la notion d'autocorrélation spatiale. C'est ce que font les analyses factorielles spatiales.
- Enfin, le lien entre diversité et marqueurs est assuré par les propriétés de dualité des analyses factorielles.

2.1 Etablir un compromis : l'analyse de coinertie multiple en génétique des populations

On veut identifier la structure consensuelle, commune à toutes les tableaux. L'analyse de coinertie multiple de Chessel and Hanafi (1996) répond à cet objectif. Cette méthode est fondée sur un critère d'optimisation de covariance : elle calcule des axes tels que les scores des observations soient à la fois dispersés (optimisation de la variance des scores), et ressemble à une structure commune, dite de référence (optimisation du carré de corrélation entre scores et

score de référence). Optimiser à la fois la variance et le carré de corrélation revient à maximiser leur produit, c'est-à-dire le carré de la covariance. Le carré de la covariance s'interprète comme une valeur typologique, quantifiant l'adéquation des structures individuelles à la référence.
Laloë et al. (2007) adaptent cette méthode aux données de génotypage où des individus regroupés en populations sont génotypés sur des microsatellites multialléliques [1]. On y propose d'abord l'utilisation d'une ACP particulière pour analyser les données d'un seul microsatellite, l'ACP sur données compositionnelles (Crespin de Billy et al., 2000), qui généralise la représentation triangulaire (du type "sable-limon-argile"). Ces représentations sont ensuite coordonnées pour en extraire la structure commune (référence) à tous les marqueurs. Enfin, des figures diagnostics permettent de visualiser l'adéquation de chaque structure individuelle à la référence (Figure 2.2).

1. Ce papier est le fruit d'une collaboration avec le professeur Chessel et un doctorant, Thibaut Jombart.

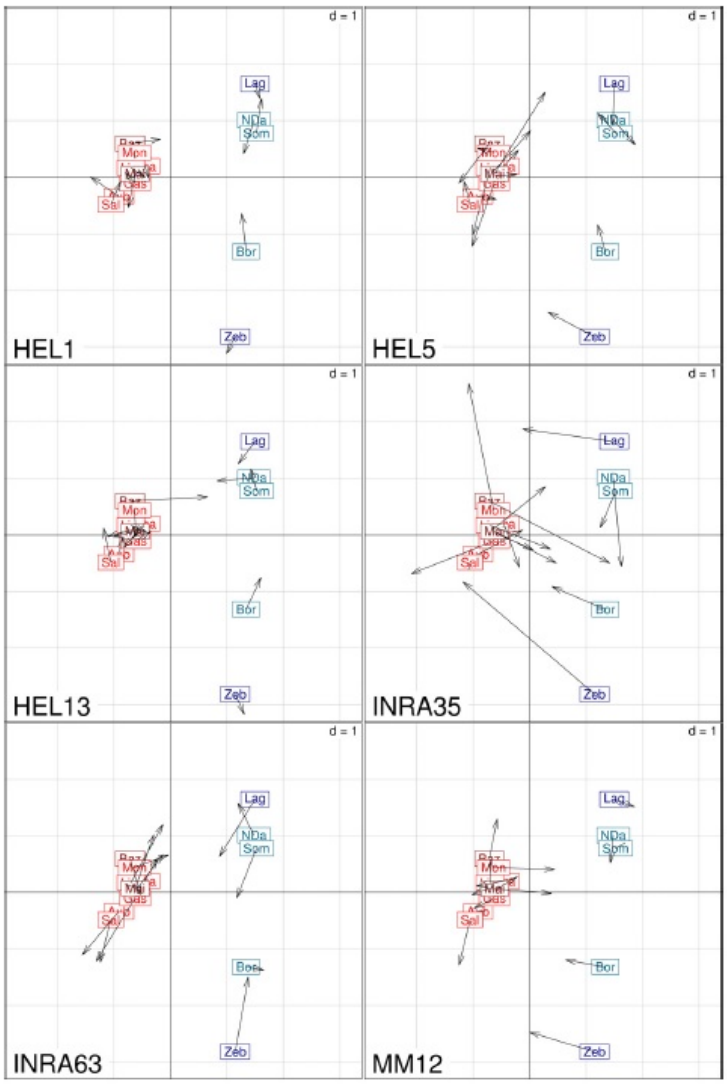


FIGURE 2.2 – Coinertie multiple. Figures montrant la différence entre la typologie de référence (étiquettes), et les analyses de chaque microsatellite, sur les deux premiers axes. La longueur des flèches visualise la difficulté d'un marqueur à reproduire la typologie de référence. La méthode est appliquée à des populations bovines (en bleu, races africaines ; en rouge, races françaises). HEL1 reproduit très bien la typologie de référence, à l'inverse de INRA35

2.2 Intégrer de données de différentes sources : l'analyse factorielle multiple sur des données zootechniques

Un autre objectif est de décortiquer, dans la "métastructure" créée par une série de plusieurs tableaux, ce qui revient à tel ou tel groupe de variables. Ainsi, l'Analyse Factorielle Multiple (Escofier and Pagès, 1998) est une analyse (ACP ou Analyse des correspondances multiples, selon le type de variables) avec une pondération particulière permettant d'équilibrer l'influence de chaque groupe. Elle fournit une structure moyenne, ou synthétique, intégrant l'ensemble des groupes, ainsi que des structures partielles, résultant des analyses réalisées groupe par groupe. Concrètement, une telle analyse consiste en :

- une ACP partielle, établie pour chaque facteur,
- une ACP pondérée par les inverses des valeurs propres maximales de chaque table : de cette façon, l'inertie axiale maximale de chaque groupe est égale à 1 ; cette ACP produit la structure synthétique ;
- une représentation simultanée de la structure moyenne et des structures de chaque table (cf Figure 2.3)
- une quantification du lien entre les groupes de variables et les axes (inertie projetée).

2.2.1 Analyse Factorielle Multiple sur les paramètres d'abattage chez le porc

J'ai appliqué cette méthode dans le cadre de la thèse de Btissam Salmi, que j'ai encadrée pour cette partie (Salmi et al., 2010)
Des porcs castrés de deux races différentes (Large White (LW) et Basque (B) sont élevés selon trois systèmes différents, intensif, alternatif et extensif, constituant cinq combinaisons race*système. Ils sont mesurés sur des variables regroupées en 9 groupes, dont des mesures de stress à l'abattage, et des profils transcriptomiques et protéomiques. On effectue une Analyse Factorielle Multiple entre les cinq combinaisons race*système d'élevage.
La figure 2.3 synthétise les résultats de l'étude : deux structures principales, la première, concrétisée par le premier axe, qui sépare les races. La deuxième individualise les systèmes d'élevage. L'influence des différents groupes de variable est visualisée de deux façons :

- les projections partielles de chaque combinaison race*système sur la figure 2.3, symbolisées par les points de différentes couleurs,
- la figure 2.4 qui donne le lien des différents groupes avec les deux premiers axes.

Ces deux figures montrent en particulier que la réactivité à l'abattage est étroitement associée aux systèmes d'élevage, et ne dépend pas de la race :

- Les points partiels correspondant aux variables "stress à l'abattage" sont très proches de l'axe 2 (Figure 2.3), la dispersion sur l'axe 1 est très faible ;
- le lien entre réactivité et le premier axe (races) est quasi-nul, alors qu'il est très élevé pour l'axe 2 (systèmes d'élevage)

Ainsi, sur une seule figure, qui, pour moi, illuste parfaitement la puissance de la méthode, on met en évidence les deux facteurs de structuration, race et système d'élevage, qui construisent les deux axes, et l'influence de chaque groupe de variables sur cette structuration.

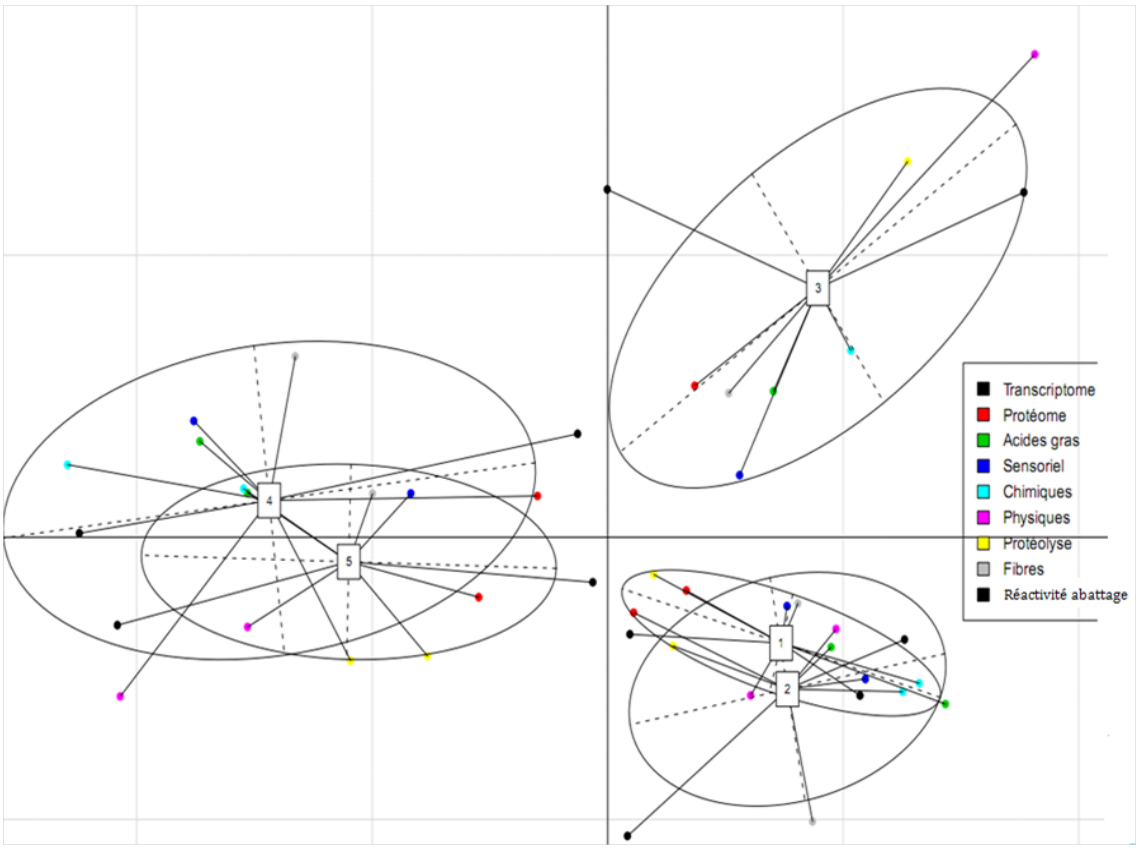


FIGURE 2.3 – Analyse Factorielle Multiple. Consensus et représentations partielles de races (Basque (B) ; Large White (LW)) et systèmes d'élevage (intensif, extérieur, extensif) selon neuf groupes de variables. 1 : B intensif ; 2 : B alternatif ; 3 : B extensif ; 4 : LW intensif ; 5 : LW alternatif.

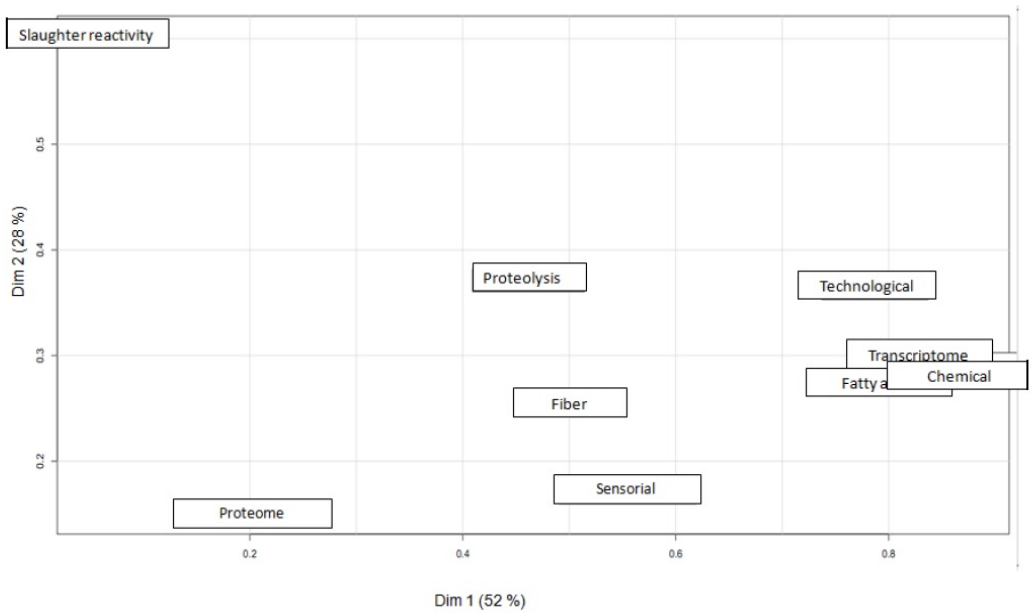


FIGURE 2.4 – Analyse Factorielle Multiple. Lien (inertie projetée) entre les groupes de variables et les deux premiers axes de l'AFM.

2.2.2 Analyse Factorielle Multiple sur les paramètres de mise-bas chez le porc

L'AFM s'utilise également dans une perspective différente, où la structuration entre groupes concerne les individus, et où l'on cherche à étudier l'influence des groupes d'individus sur les relations existant entre les variables. Ainsi, durant le doctorat de Laurianne Canario, nous avons caractérisé la structure des relations entre les paramètres de la mise-bas chez le porc, et la variation de cette structure selon la race. Canario et al. (2009) ont montré que les relations entre variables, visualisées par un cercle de corrélation (Figure 2.5) sont stables quelles que soient les races, même si la Meishan s'écarte un peu de la structure commune, en particulier pour les variables d'homogénéité de poids.

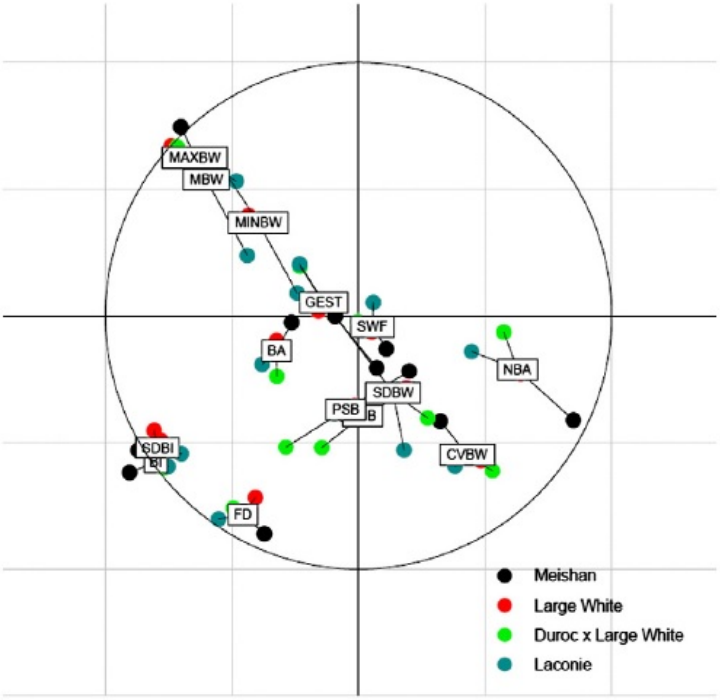


FIGURE 2.5 – Analyse Factorielle Multiple. Relations entre prolificité, reproduction et croissance chez le porc en fonction de la race. Structures globale (label) et raciales (points colorés). Les variables incluent le nombre total de porcelets nés (NBT), nés vivants (NBA), mort-nés (NSB) par portée ; le pourcentage de mort-nés (PSB), les durées de gestation (GEST) et de mise-bas (FD) ; l'intervalle entre naissances (BI) et son écart type (SBI), l'aide à la mise-bas(BA) ; pour le poids à la naissance, la moyenne (MBW), le coefficient de variation (CBW), l'écart type (SDBW), le maximum (MAXBW), le minimum (MINBW) et le poids de la truie à la mise-bas (SWF)

2.3 Interpréter la diversité génétique par la géographie : ACP spatiale

L'analyse des structures spatiales repose sur l'analyse de l'autocorrélation spatiale (Sokal and Wartenberg, 1983). Les analyses factorielles spatiales ont été proposées par Thioulouse et al. (1995) et adaptées au contexte spécifique de la génétique des populations par Jombart et al. (2008). L'analyse se concentre sur la part de variance structurée spatialement, en incorporant l'information spatiale dans le critère à optimiser. On juge ensuite de l'importance de cette structuration par une comparaison avec les résultats d'une ACP classique. Les structures spatiales se révèlent sur une carte géographique, où chaque observation est symbolisée par un point dont le code couleur combine jusqu'à trois scores via le système RVB.
L'ACP spatiale analyse une matrice de données **X**, conjointement avec une information spatiale contenue dans une matrice **L**, dérivant le plus souvent d'un graphe de voisinage connectant les populations voisines sur une carte. **L** est utilisée pour le calcul de l'autocorrélation spatiale (Indice de Moran) d'une variable centrée **x** : $I = \frac{\mathbf{x'Lx}}{\mathbf{x'x}}$. Cet indice est positif quand les voisins sont semblables (structure globale), et négatif quand les voisins sont différents (structure locale). L'ACP spatiale effectue la décomposition canonique de la matrice **X'(L+L')X**. Les valeurs propres peuvent être positives ou négatives, selon la présence de structures globales ou locales. Des test permettent d'apprécier la significativité des structures décelées.
J'ai principalement utilisé cette méthode dans deux études.

2.3.1 Structure spatiale de la diversité génétique des ruminants d'Europe et d'Asie

Laloë et al. (2010) étudient la structuation spatiale des ruminants (Bovins, caprins, ovins) d'Europe et d'Asie, génotypés sur des microsatellites dans le cadre de deux projets européens. A titre d'exemple, 45 populations caprines ont été typés sur 30 microsatellites. Les principaux résultats de l'ACP spatiale sont visualisés sur la figure 2.6, pour les cinq premiers axes. Les deux premiers axes montrent un gradient Sud-Est *vs* Nord-Ouest pour le premier, Sud *vs* Nord pour le deuxième, isolant les populations ibériques. Les autres axes montrent des structures spatiales indécelables sur l'ACP classique, et concernant des regroupements de populations à échelle plus fine.

2.3.2 Structure spatiale de la diversité génétique des bovins français

Gautier et al. (2010) recherchent les patrons spatiaux de la diversité génétique entre 23 races bovines européennes, principalement françaises. On observe (Figure 2.7) une homogénéité génétique importante pour les races des montagnes de l'Est, au contraire du Grand Ouest, très hétérogène. Cette hétérogénéité s'explique par des phénomènes de croisement et d'introgression bien documentés, en particulier pour les races Maine-Anjou (MAN) et Pie Rouge des Plaines (PRP).

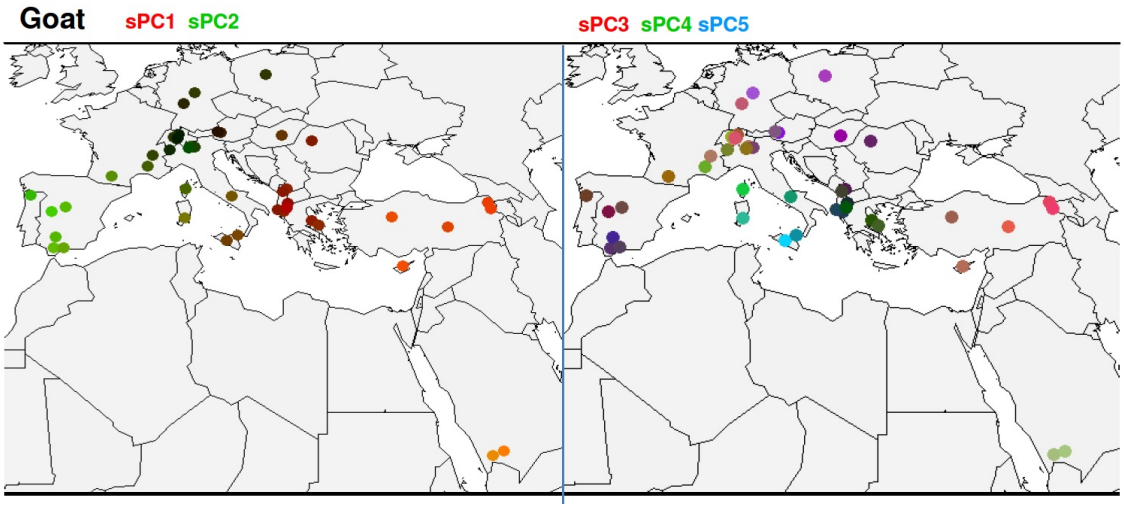


FIGURE 2.6 – ACP spatiale. Géographie et diversité génétique des caprins - Axes 1 à 5

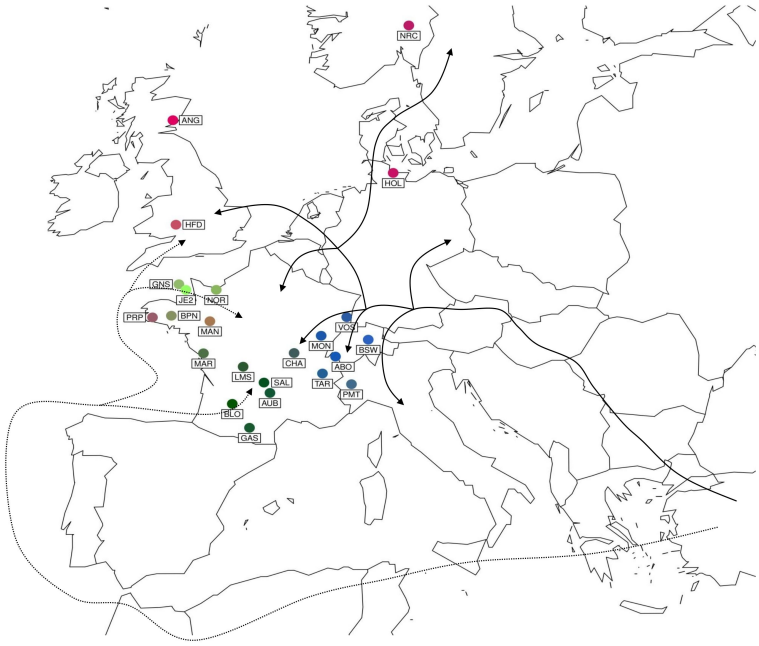


FIGURE 2.7 – ACP spatiale. Géographie et diversité génétique des bovins français - Axes 1 et 2. Les routes de migration sont représentées par des lignes pleine (Danubienne) et pointillée (Méditerranéenne)

2.3.3 Interprétation populationnelle de la diversité génétique

Il peut paraître curieux de consacrer une section à l'interprétation de la diversité génétique en termes populationnels, tant le regroupement d'individus en populations est naturel, en tout cas pour les animaux domestiques. Néanmoins, en génétique, l' "atome" est l'haplotype, c'est-à-dire la forme allélique d'un marqueur sur un chromosome de l'individu. Toute autre analyse porte sur des synthèses, que ce soit le génotype pour les espèces diploïdes, ou la population. L'interprétation génétique de ces différentes analyses synthétiques a été étudiée par Laloë and Gautier (2011), dans le cadre de SNP bialléliques, qui sont devenus aujourd'hui le cadre de référence des analyses de population. On peut appliquer aux données d'haplotype f_{ijk}, correspondant au *k*ième haplotype du *j*ème individu de la *i*ème population, la décomposition suivante :

$$\begin{aligned} f_{ijk} - f_{...} = f_{ijk} - f_{ij.} \\ + f_{ij.} - f_{i..} \\ + f_{i..} - f_{...} \end{aligned}$$

On a les résultats suivants :

- A chaque terme de la décomposition correspond une ACP. En particulier, au terme $f_{i..} - f_{...}$ correspond une ACP entre races.
- Si les haplotypes sont normés par $\sqrt{f_{...}(1 - f_{...})}$, il y a équivalence entre ACP normée, analyse des correspondances multipes et ACP sur la matrice des corrélations *r*, où *r* est la racine carrée de la mesure du déséquilibre de liaison entre marqueurs.
- Les F-statistiques peuvent s'estimer en fonction des inerties des différentes ACP. Ainsi, *Fst* s'estime par le rapport d'inertie de l'ACP entre races et l'inertie totale (égale au nombre de SNP).
- Le carré de la coordonnée d'un SNP sur une composante est le R^2 du modèle linéaire liant coordonnée de l'haplotype sur l'axe j et la forme allélique du marqueur

  $$y = \mu + SNP + e$$

  (cf Figure 2.8). On l'interprétera comme une valeur typologique, c'est-à-dire comme le potentiel du SNP à reproduire la typologie décrite par l'axe correspondant.

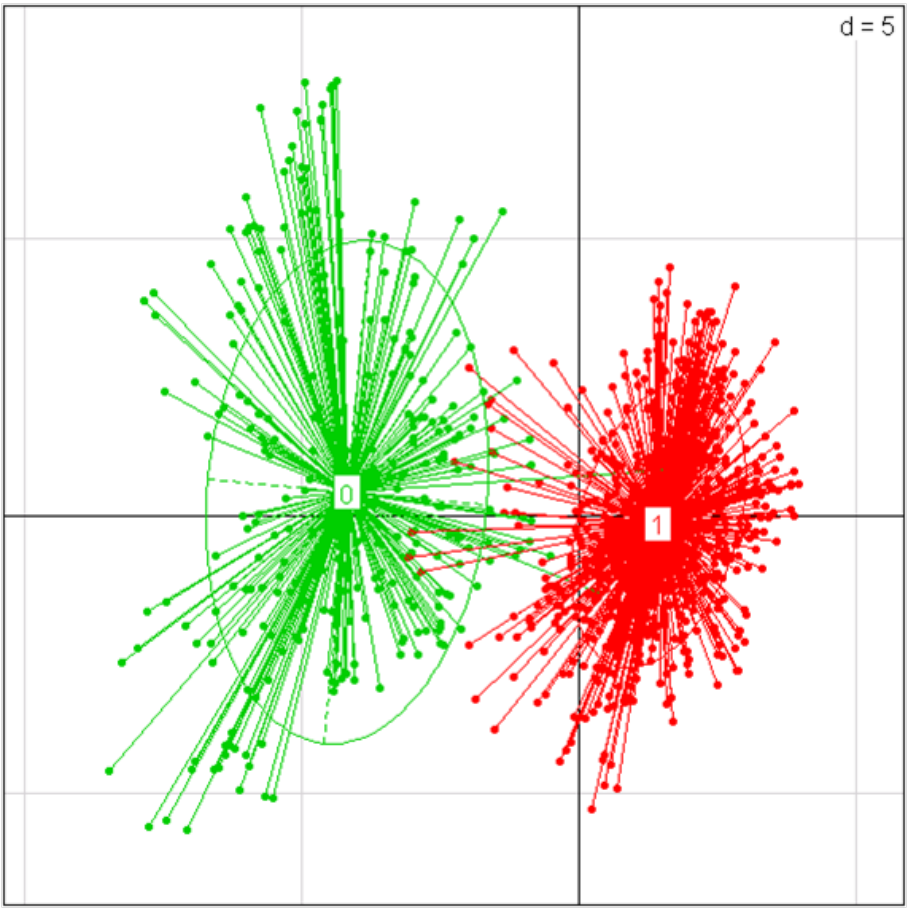


FIGURE 2.8 – Les deux premiers axes d'une ACP normée sur haplotypes. En vert, les individus porteurs de forme "0" d'un SNP donné ; en rouge, les individus porteurs de la forme "1" du même SNP. Les valeurs typologiques (carrés des coordonnées) du SNP sont égales à 0,79 pour la première composante et à 0,01 pour la deuxième.

En résumé :

- La diversité génétique peut s'interpréter à l'aide de variables exogènes. L'ACP entre populations en est un exemple simple, qui s'étend facilement à tout modèle linéaire. Ce sont les ACP sur variables instrumentales (Rao, 1964)
- A toute typologie révélée par une ACP, correspondent des valeurs typologiques de SNP, dont l'étude permettra de dégager des régions génomiques associées à la diversité génétique. On associera une interprétation de la diversité génétique par des variables exogènes (environnement, climat,...) à des régions génomiques. C'est l'objet de la génomique environnementale.

2.4 Interpréter la diversité génétique par l'environnement (Génomique environnementale)

La génomique environnementale naît de la rencontre de deux grands phénomènes :

- Le réchauffement climatique, et l'urgence de considérer l'adaptation des êtres vivants à un environnement changeant, sous ses différentes formes (robustesse, résilience, résistance à la chaleur et à la sécheresse,...) ;
- Le développement des NGS (*Next Generation Sequencing*), qui permettent de cerner au plus près les réalités topologique et structurelle du génome, et d'identifier in fine les régions génomiques que l'on peut associer à l'adaptation.

Nous avons préconisé, dans deux interventions orales de congrès, l'utilisation des méthodes factorielles pour atteindre ce but.

2.4.1 Les analyses sur variables instrumentales

Laloë and Zerjal (2014) détaillent l'intérêt de l'ACP sur variables instrumentales (ACPVI) intégrant variables climatiques et données génétiques. L'ACPVI modélise les relations entre le tableau X et les variables explicatives (ou instrumentales) via un modèle linéaire. C'est une extension de l'ACP d'un tableau $\mathbf{X}$ modélisé par le tableau $\mathbf{W}$, incluant des variables quantitatives ou qualitatives. L'analyse se fait sur les variables prédites selon le modèle $\mathbf{X} = \mathbf{W} + \mathbf{e}$, et donc sur $\hat{\mathbf{X}}'\hat{\mathbf{X}}$, avec $\hat{\mathbf{X}} = \mathbf{W}[\mathbf{W}'\mathbf{W}]^{-}\mathbf{W}'\mathbf{X}$. Les composantes issues de cette analyse appartiennent au sous-espace engendré par $\mathbf{W}$, et maximisent la somme des carrés des corrélations avec les variables de $\mathbf{X}$. Le critère à optimiser combine donc variabilité et "prédictabilité". Cette analyse permet de plus de mesurer la variance expliquée par différents ensembles de variables, et de tester séquentiellement la signification des variables par un test de permutation (Legendre and Legendre, 2012)

2.4.2 Détecter les signaux de différenciation : le Fused Lasso

Une ACPVI permet de dégager les axes de diversité génétique covariantes avec l'environnement et les valeurs typologiques correspondantes des SNP. Les régions qui nous intéressent

concentrent le plus de SNP à haute valeurs typologiques. C'est l'objectif du Fused Lasso (Tibshirani et al., 2005), présenté dans ce contexte par Laloë et al. (2014). Le Fused Lasso cherche à isoler, dans un vecteur ordonné, les points consécutifs avec des valeurs élevées et constantes (Figure 2.9), les résultats pouvant se visualiser à l'aide d'un "Manhattan Plot" (Figure 2.10). La difficulté de ce travail en cours, conduit en collaboration avec Julien Chiquet (AgroParisTech, Génopole d'Evry), Mathieu Gautier (INRA CBGP-Montpellier) et Florence Jaffrézic (INRA - GABI), consiste à trouver le meilleur choix du couple de paramètres (λ_1, λ_2) contrôlant la "sparsité" et l'aspect spatial. Plusieurs stratégies sont actuellement à l'étude, fondées sur des critères de pénalisation (*i.e.* AIC, BIC) ou des procédures de "stability selection" (Meinshausen and Bühlmann, 2010).

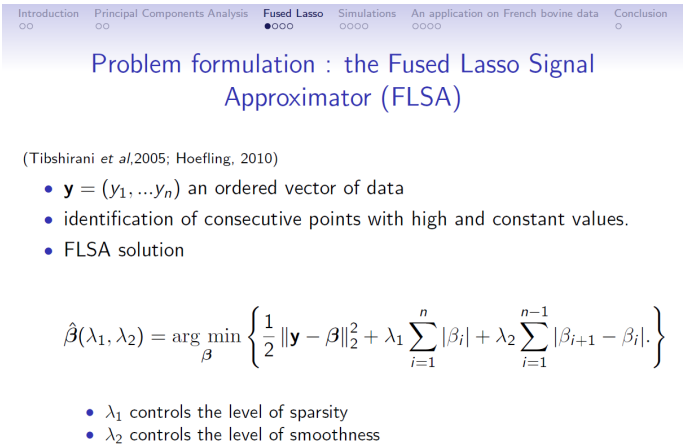


FIGURE 2.9 – Présentation du Fused Lasso (Laloë et al., 2014)

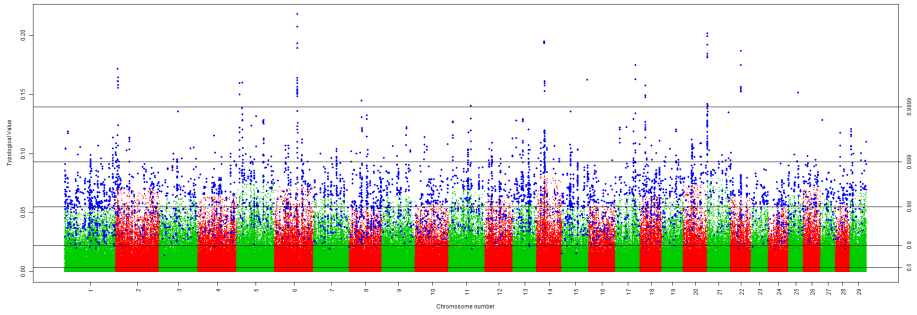


FIGURE 2.10 – Une application du Fused Lasso à la détection de régions génomiques différenciant les bovins laitiers des bovins allaitants, à partir de génotypes réalisés sur une puce Haute Densité. Valeurs typologiques de SNP selon leur localisation sur le génome ; en bleu, régions identifiées par la procédure

2.4.3 Une application : Le projet GALIMED, ou l'adaptation des bovins méditerranéens aux contraintes climatiques.

Nous appliquerons ces méthodes aux données générées dans le cadre d'un projet que j'ai coordonné. Ce projet, dont l'acronyme est *GALIMED* [2] a été financé par ACCAF, un métaprogramme de l'Inra [3]. Il s'appuie sur les prévisions concernant le changement climatique dans les régions méditerranéennes (augmentation des températures, diminution des précipitations, extension du climat méditerranéen vers le nord, en particulier le Massif Central). En conséquence, l' aptitude du bétail à s'adapter aux variations climatiques va devenir un facteur de première importance. Ce projet cherche à comprendre les bases génétiques de l'adaptation aux contraintes climatiques, via une approche multidisciplinaire combinant génétique des populations, description de l'environnement et des systèmes de production. On étudie des races bovines locales, situées sur les deux rives de la Méditerranée (Algérie, Chypre, Egypte, Espagne, France, Grèce, Italie, Maroc). Comme ces races sont adaptées depuis des siècles à leur environnement, l'étude de leur diversité en fonction des paramètres climatiques et des systèmes d'élevage permettra de détecter des empreintes de sélection associées à l'adaptation à la chaleur et à la sécheresse (cf Figure 2.11)

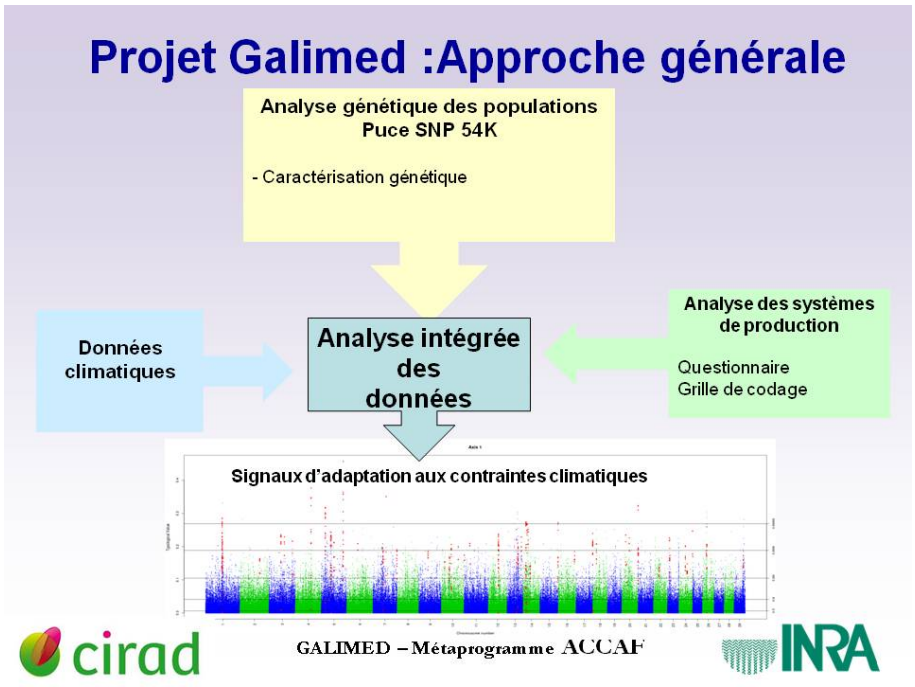


FIGURE 2.11 – Le projet Galimed

2. Génétique de l'adaptation des bovins et les Systèmes de production en Méditerranée
3. Adaptation de l'Agriculture et de la Forêt au Changement Climatique

2.5 Promotion des analyses factorielles

Les analyses factorielles sont souvent présentées d'une façon réductrice, ne serait-ce que par le vocabulaire qui les désigne : *descriptive, exploratoire, résumé,...*. La démarche descriptive-inductive (Le Roux, 2014) qui sous-tend ces approches est à rebours de la démarche classique inférentielle, sans parler de l'accent mis sur les graphiques, perçus comme non scientifiques. Il y a encore des revues exigeant de présenter la table de corrélations *in extenso* plutôt que le cercle des corrélations (cf Figure 3). En génétique des populations, l'article le plus cité sur l'ACP (Patterson et al., 2006) la réduit à une simple opération matricielle (*Eigenanalysis*). C'est pourquoi j'essaye de promouvoir ces analyses devant des auditoires variés (biologistes, généticiens, étudiants), de pays différents (cf 2.12), en mettant l'accent sur les points suivants :

- Visualisation synthétique d'un phénomène, avec la possibilité de partitionner les variables en plusieurs groupes, ce qui ouvre la voie à l'intégration de données hétérogènes ;
- Dualité des points de vue, entre observations et variables ;
- Visualisation des structures de populations, en utilisant des marqueurs dont le nombre varie de quelques dizaines (microsatellites) à plusieurs centaines de milliers (puce Haute Densité), et ce, en utilisant le logiciel R et ses packages, pour des durées de traitement n'excédant pas quelques heures ;
- Interprétation de la diversité génétique en fonction de la géographie ou du regroupement des individus en population ;
- Liens entre ratios d'inertie et paramètres génétiques (*Fst*)
- Etablissement de typologies, consensuelles ou liées à des variables exogènes, et contribution de chaque marqueur à la construction de cette typologie.

FIGURE 2.12 – Quelques présentations des analyses factorielles

3

Perspectives

Les directions générales selon lesquelles je compte orienter mes activités de recherche, et dans lesquelles l'analyse géométrique des données tiennent leur place, s'appuient sur trois considérations :

- Le déferlement de données massives et hétérogènes ;
- changement de nature des informations génétiques (NGS) ;
- décloisonnement de la génétique quantitative.

.

3.1 Données massives et hétérogènes. Biologie intégrative et analyses factorielles.

Dans un contexte de génération de données massives, on peut pronostiquer sans trop de risque un usage croissant des analyses factorielles. Les données zootechniques classiques (performances de production, de fertilité...), obtenues au niveau de l'animal, ont peu à peu été, sinon remplacées, en tout cas complétées par toute une série de données biologiques fines, généralement suffixées *.-omiques* : transcriptomique, génomique, métabolomique, hétérogènes par leur nature et leur interprétation. Une étude typique s'appuie sur l'analyse d'un phénomène biologique dans ses multiples composantes, complétée par l'examen de l'impact d'un facteur (pollution par des nanoparticules, pratique d'élevage, type d'alimentation) sur l'ensemble de ces données. Je suis partie prenante dans plusieurs projets de ce type. La structure du centre INRA de Jouy-en-Josas regroupant plusieurs laboratoires générant des données hétérogènes et sollicitant leur analyse est un atout en ce sens.

3.2 Changement de nature des informations génétiques. Une thèse sur le concept de race à l'ère génomique

Je compte déposer un sujet de thèse sur le concept de thèse à l'ère génomique, à partir de l'argument suivant :
La modification de la nature des informations génétiques et la connaissance de plus en plus fine du génome ont bouleversé la façon d'appréhender les mécanismes génétiques d'expression des caractères, et l'évaluation génétique. Parallèlement, elles devraient également changer la notion de race, centrale en amélioration génétique, est principalement fondée sur des critères phénotypiques et de suivi généalogique, avec le principe du standard introduit au XIXème siècle en Europe et la création de livres généalogiques. La race est également, sinon davantage, un concept culturel dont la définition varie selon les régions du monde. En première analyse, une race est constituée d'animaux homogènes, sans structuration particulière. Il s'agira donc de proposer, à partir de données génomiques (SNP, séquences) cette absence de structure dans un groupe d'animaux censés constituer une race. Cette problématique reprend de l'importance dans une optique d'évaluation et de conservation de la biodiversité, dans des contextes différents :

- La reconstitution d'une race. Certaines races, quasi-éteintes, sont reconstituées à l'aide de croisements entre animaux préservés et animaux de races proches, sur la base d'un standard phénotypique. C'est par exemple le cas d'une race belge flamande, la Kempisch Rund.
- La caractérisation de races locales, particulièrement dans les pays en développement, où cette notion ne repose pas sur des standards précis ou des livres généalogiques. Ainsi, au Maghreb (Algérie et Maroc), où une analyse en composantes principales de populations maghrébines et européennes, distingue trois types de population du Maghreb, selon leur degré de préservation vis-à-vis de races européennes (cf Figure 3.1) :
  - des populations locales "préservées" de l'influence européenne : les races Tidili, Oulmès et Guelmoise ;
  - des populations "contaminées" par des croisements avec des races européennes : c'est le cas de la Chélifienne, population dans laquelle coexistent des individus locaux et croisés ;
  - des populations issues de croisements entre races européennes : la Biskra.

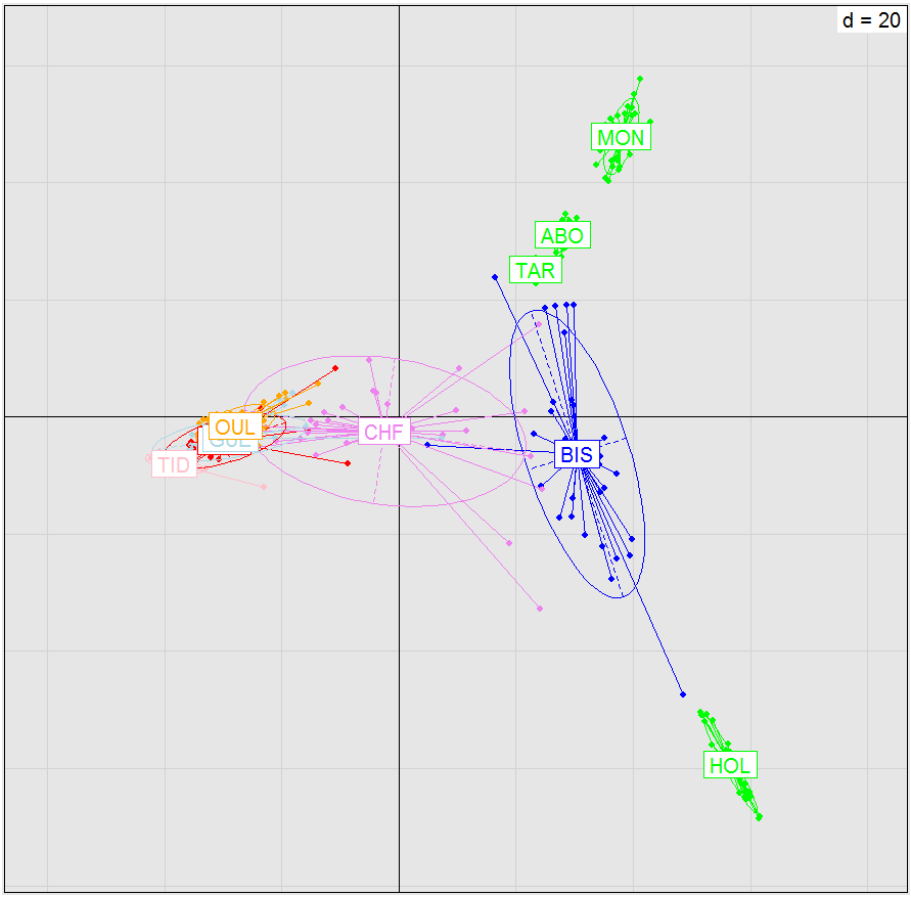


FIGURE 3.1 – ACP sur races maghrébines et européennes. En vert, races européennes : HOL (Holstein), MON (Montbéliarde), ABO (Abondance), TAR (Tarine) ; les autres couleurs représentent les populations du Maghreb : BIS (Biskra), CHF (Chélifienne), GUE (Guelmoise), OUL (Oulmès), TID (Tidili)

3.3 Décloisonner la génétique quantitative des espèces domestiques

La génétique des espèces domestiques est une discipline méconnue et isolée. L'étude sur les fondements de l'évaluation génétique que j'ai réalisée (Laloë, 2011) m'a fait prendre conscience à quel point la problématique liée à l'évaluation génétique avaient été déterminante pour établir les concepts fondamentaux de la génétique (cf section 1.1.1). D'autre part, les populations d'animaux domestiques sont un cas d'école pour l'étude génétique des populations, grâce à la connaissance que l'on peut avoir de leur structure, à travers notamment les généalogies, mais également l'histoire. Enfin, ces populations disposent aujourd'hui de polymorphismes moléculaires, obtenues à partir de puces spécifiques, ou de séquences. Pourtant, peu de rapports existent entre généticiens des populations domestiques et sauvages.

3.3.1 Méconnaissance de la génétique quantitative. Un post-doctorat sur l'histoire de la génétique quantitative

J'ai proposé à Thomas Heams, enseignant-chercheur de l'équipe PSGen et responsable de l'UC "Epistémologie et Histoire des Sciences" à AgroParisTech, de monter un projet de post-doctorat sur ce thème.

3.3.2 Décloisonner la génétique des populations domestiques. Une thèse sur les variables essentielles de biodiversité

Nous allons reproposer un sujet de thèse sur l'argument suivant :
Les variables essentielles de biodiversité (VEB) sont des mesures destinées à l'étude, au suivi et à la gestion de la biodiversité. Plusieurs mesures ont été proposées pour l'analyse de la composition génétique des populations par le Groupe d'observations de la Terre pour l'observation de la biodiversité (GEO-BON). Ces mesures sont destinées à caractériser la diversité génétique intra-spécifique, pour des populations domestiques ou sauvages. Les races animales domestiques disposent d'abondantes données de généalogie et, plus récemment, de polymorphismes moléculaires, qui permettent de renseigner les VEB et de tester leur sensibilité au type de données collectées d'une part, à des changements de gestion d'autre part. La situation est bien différente pour les populations sauvages où les données permettant de renseigner les VEB "composition génétique" sont beaucoup plus rares. Le premier objectif de la thèse est d'évaluer les VEB "composition génétique" sur une gamme de populations animales sélectionnées ou en conservation, différant par leur intervalle de génération (bovins, chevaux, poulets) et par la quantité d'informations disponibles. Le second objectif est d'étudier les conditions de l'application de ces VEB à des populations sauvages, concernant le type de données à recueillir et les méthodes d'analyse à utiliser.La structuration des populations sera étudiée via des méthodes factorielles car elles peuvent être reliées aux F-statistiques et mises en oeuvre de manière non supervisées ou en relation avec des caractéristiques environnementales ou géographiques.

3.3.3 Décloisonner la génétique des populations domestiques. Formation et réseautage

Nous projetons au sein de l'équipe PSGen, de développer un réseau regroupant généticiens des populations domestiques et sauvages, autour de thèmes communs (diversités des populations, taille efficace,...). Grâce à la présence dans l'équipe d'enseignants-chercheurs d'AgroParisTech, nous disposons de savoir-faire et de matériel pédagogique innovant (MOOC, Podcasts) nous permettant d'envisager à court terme le montage d'un module de formation doctorale, sur le thème de la détection de signaux d'adaptation à l'environnement assuré conjointement par des généticiens des populations domestiques et sauvages.

Conclusion

Je n'épiloguerai pas sur l'intérêt d'utiliser des méthodes géométriques pour analyser les données génétiques. Ce mémoire en présente suffisamment d'exemples.
Je préfère m'attarder, au risque d'être ridicule et de pulvériser mon seuil de compétence, sur des aspects plus philosophiques et personnels.
Je me suis souvent interrogé sur l'origine de l'inclination que j'éprouve pour ces méthodes, en particulier les méthodes factorielles, et j'en suis venu à les juger à l'aune de la philosophie matérialiste d'Epicure que j'apprécie particulièrement. Les affinités sont nombreuses :

- Les méthodes factorielles sont fondamentalement empiriques et se fondent sur l'observation, et l'induction ;
- elles privilégient les graphiques et donc une approche sensualiste ;
- il faut connaître l'ensemble avant de s'intéresser aux détails ;
- comment ne pas faire l'analogie entre les points de poussière volant au soleil, évoqués par Lucrèce pour justifier l'hypothèse atomiste, et les cartes factorielles ?
- Les méthodes factorielles se bornent (ou en tout cas, elles devraient se borner) à interpréter des structures révélées par des graphiques, sans en tirer des conclusions hâtives : toute affirmation non fondée sur les sens relève de la spéculation.

Enfin, pour Epicure, la science a pour but de délivrer les hommes des craintes dues à l'ignorance, et est subordonnée à la recherche du plaisir et du bien-être de l'homme. Les généticiens l'ont oublié parfois. C'est en tout cas un belle motivation pour faire de la recherche scientifique.

Bibliographie

S. Andonov, G. Klemetsdal, and T. Adnoy. Computer intensive calculations of reliability of predicted breeding values with animal model, as a mean for decision making in practical breeding schemes. In *ID881 in Proceedings of the 10th World Congress Genetics Applied to Livestock Production, Vancouver, Canada*, 2014.

M. Armatte. Histoire et préhistoire de l'analyse des données par JP Benzecri : un cas de généalogie rétrospective. *Journal électronique d'histoire des probabilités et de la statistique*, 2008.

L. Canario, Y. Billon, J. C. Caritez, J. P. Bidanel, and D. Laloë. Comparison of sow farrowing characteristics between a chinese breed and three french breeds. *Livestock Science*, 125(2) : 132–140, 2009.

L. L. Cavalli-Sforza. Population structure and human evolution. *Proceedings of the Royal Society of London. Series B*, 164 :362–379, 1966. 00297.

D. Chessel and M. Hanafi. Analyses de la co-inertie de k nuages de points. *Revue de statistique appliquée*, 44(2) :35–60, 1996.

J. Coursol. *Technique statistique des modèles linéaires*. Les cours du CIMPA. CIMPA, Nice, 1980.

V. Crespin de Billy, S. Doledec, and D. Chessel. Biplot presentation of diet composition data : an alternative for fish stomach contents analysis. *Journal of Fish Biology*, 56(4) :961–973, 2000.

D. Cros, M. Denis, L. Sanchez, B. Cochard, A. Flori, T. Durand-Gasselin, B. Nouy, A. Omoré, V. Pomiès, V. Riou, E. Suryana, and J.-M. Bouvet. Genomic selection prediction accuracy in a perennial crop : case study of oil palm (elaeis guineensis jacq.). *Theoretical and Applied Genetics*, pages 1–14, 2014.

V. Dodelin, F. Phocas, and D. Laloë. Robustness of an animal design. In *Proceedings of the 51st Annual Meeting of the European Association for Animal Production*, The Hague, The Netherlands, 2000.

S. Dray and A. Dufour. The ade4 package : implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22(4) :1–20, 2007.

C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3) :211–218, 1936.

B. Escofier and J. Pagès. *Analyses factorielles simples et multiples. Objectifs, méthodes et interprétation*. Dunod, Paris, 1998.

Y. Escoufier. The Duality Diagram : A Means for Better Practical Applications. In P. Legendre and L. Legendre, editors, *Develoments in Numerical Ecology*, pages 139–156. Springer Berlin Heidelberg, Berlin, Heidelberg, 1987.

J. Felsenstein. Confidence Limits on Phylogenies : An Approach Using the Bootstrap. *Evolution*, 39(4) :783, 1985.

R. A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Philosophical Transactions of the Royal Society of Edinburgh*, 52 :399–433, 1918.

M. N. Fouilloux and D. Laloë. A sampling method for estimating the accuracy of predicted breeding values in genetic evaluation. *Genet. Sel. Evol*, 33 :473–486, 2001.

M.-N. Fouilloux, V. Clement, and D. Laloë. Measuring connectedness among herds in mixed linear models : From theory to practice in large-sized genetic evaluations. *Genetics Selection Evolution*, 40(2) :145–159, 2008a.

M. N. Fouilloux, S. Minery, S. Mattalia, and D. Laloë. Assessment of connectedness in the international genetic evaluation of simmental and montbeliard breeds. *Bulletin*, 35 :129–135, 2008b.

J. L. Foulley, J. Bouix, B. Goffinet, and J. M. Elsen. Connectedness in genetic evaluation. In P. D. D. Gianola and D. K. Hammond, editors, *Advances in Statistical Methods for Genetic Improvement of Livestock*, number 18 in Advanced Series in Agricultural Sciences, pages 277–308. Springer Berlin Heidelberg, Jan. 1990.

M. Gautier, D. Laloë, and K. Moazami-Goudarzi. Insights into the genetic history of french cattle from dense SNP data on 47 worldwide breeds. *PLoS One*, 5(9) :e13038, 2010.

C. R. Henderson. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31(2) :423–447, 1975.

C. R. Henderson. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics*, 32(1) :69–83, 1976.

D. G. Herr. On the history of the use of geometry in the general linear model. *The American Statistician*, 34(1) :43–47, 1980.

J. Isidro, J.-L. Jannink, D. Akdemir, J. Poland, N. Heslot, and M. E. Sorrells. Training set optimization under population structure in genomic selection. *Theoretical and Applied Genetics*, 2014.

T. Jombart, S. Devillard, A. B. Dufour, and D. Pontier. Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity*, 101(1) :92–103, 2008.

B. Kennedy and D. Trus. Considerations on genetic connectedness between management units under an animal model. *Journal of animal science*, 71(6) :2341–2352, 1993.

L. A. Kuehn, R. M. Lewis, and D. R. Notter. Managing the risk of comparing estimated breeding values across flocks or herds through connectedness : a review and application. *Genetics, Selection, Evolution*, 39(3), 2007.

D. Laloë. Precision and information in linear models of genetic evaluation. *Genetics Selection Evolution*, 25(6) :557–576, 1993.

D. Laloë. La genèse et le développement des concepts de l'évaluation génétique classique. *Productions Animales*, 24(4) :323, 2011.

D. Laloë and M. Gautier. On the genetic interpretation of between-group PCA on SNP data. Technical Report hal-00661214, INRA, 2011. URL https://hal.archives-ouvertes.fr/hal-00661214.

D. Laloë and F. Phocas. A proposal of criteria of robustness analysis in genetic evaluation. *Livestock Production Science*, 80(3) :241–256, 2003.

D. Laloë and T. Zerjal. Landscape genomics and multivariate analysis : examples and prospects for poultry. In *Proceedings of the 8th European Poultry Genetics Symposium*, Venizia, Italie, 2014. URL https://hal.archives-ouvertes.fr/hal-00919421/.

D. Laloë, F. Phocas, and F. Ménissier. Considerations on measures of precision and connectedness in mixed linear models of genetic evaluation. *Genetics selection evolution*, 28(4) : 359–378, 1996.

D. Laloë, T. Jombart, A. Dufour, and K. Moazami-Goudarzi. Consensus genetic structuring and typological value of markers using multiple co-inertia analysis. *Genetics Selection Evolution*, 39(5) :545–567, 2007.

D. Laloë, K. Moazami-Goudarzi, J. A. Lenstra, P. A. Marsan, P. Azor, R. Baumung, D. G. Bradley, M. W. Bruford, J. Canon, and G. Dolf. Spatial trends of genetic variation of domestic ruminants in europe. *Diversity*, 2(6) :932–945, 2010.

D. Laloë, J. Chiquet, F. Jaffrézic, and M. Gautier. FLPCA : A Fused Lasso PCA-based approach to identify influential markers in differentiated populations from dense SNP data. In *International Biometric Conference*, Firenze, Italy, July 2014. International Biometric Society. URL https://hal.archives-ouvertes.fr/hal-01075342.

B. Le Roux. *Analyse gométrique des données multidimensionnelles*. Psycho Sup. Dunod, Paris, France, 2014.

H. Leclerc. *Mise en place de l évaluation génétique sur les controles élémentaires en bovins laitiers et perspectives d utilisation des résultats en appui technique*. Theses, AgroParisTech, Dec. 2008. URL https://pastel.archives-ouvertes.fr/pastel-00004976.

P. Legendre and L. Legendre. *Numerical Ecology*. Elsevier Science Publishers, Amsterdam, 2012.

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 72(4) :417–473, 2010.

T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4) :1819–1829, 2001.

K. Moazami-Goudarzi and D. Laloë. Is a multivariate consensus representation of genetic relationships among populations always meaningful ? *Genetics*, 162(1) :473–484, 2002.

K. Moazami-Goudarzi, J. P. Furet, F. Grosclaude, and D. Laloë. Analysis of genetic relationships between 10 cattle breeds with 17 microsatellites. *Animal Genetics*, 28(5) :338–345, 1997.

J. Novembre and A. Di Rienzo. Spatial patterns of variation due to natural selection in humans. *Nature Reviews Genetics*, 10(11) :745–755, 2009.

N. Patterson, A. L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Genet*, 2(12) :e190, 2006.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2 :559–572, 1901.

N. T. Pegolo, D. Laloë, H. N. de Oliveira, R. B. Lobo, and M.-N. Fouilloux. Trends of the genetic connectedness measures among nelore beef cattle herds. *Journal of Animal Breeding and Genetics*, 129(1) :20–29, 2012.

C. R. Rao. The use and interpretation of principal component analysis in applied research. *Sankhyā : The Indian Journal of Statistics, Series A*, pages 329–358, 1964.

R. Rincent, D. Laloë, S. Nicolas, T. Altmann, D. Brunel, P. Revilla, V. M. Rodriguez, J. Moreno-Gonzalez, A. Melchinger, E. Bauer, C.-C. Schoen, N. Meyer, C. Giauffret, C. Bauland, P. Jamin, J. Laborde, H. Monod, P. Flament, A. Charcosset, and L. Moreau. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals : Comparison of methods in two diverse groups of maize inbreds (zea mays l.). *Genetics*, 192(2) :715–728, 2012.

B. Salmi, C. Larzul, M. Damon, L. Lefaucheur, J. Mourot, E. Laville, P. Gatellier, K. Méteau, D. Laloë, and B. Lebret. Multivariate analysis to compare pig meat quality traits according to breed and rearing system. In *ID442 in Proceedings of the 9th World Congress Genetics Applied to Livestock Production, Leipzig, Germany*, 2010.

H. Scheffé. A method for judging all contrasts in the analysis of variance. *Biometrika*, 40 (1/2) :87–104, 1953.

H. Scheffé. *The Analysis of Variance*. John Wiley & Sons, 1959.

S. D. Silvey and D. M. Titterington. A geometric approach to optimal design theory. *Biometrika*, 60(1) :21–32, 1973.

R. R. Sokal and D. E. Wartenberg. A test of spatial autocorrelation analysis using an isolation-by-distance model. *Genetics*, 105(1) :219–237, 1983.

J. Tarres, M. Fina, and J. Piedrafita. Connectedness among herds of beef cattle bred under natural service. *Genetics Selection Evolution*, 42(1) :6, 2010.

J. Thioulouse, D. Chessel, and S. Champely. Multivariate analysis of spatial patterns : a unified approach to local and global structures. *Environmental and Ecological Statistics*, 2 (1) :1–14, 1995.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67 (1) :91–108, 2005.